Estratégia Concursos

Aula 07 (Prof. Thiago Cavalcanti)

IFCE (Prof. Ciência da Computação-Metodologia e Técnicas da Comput) Conhecimentos - 2021(Pós-Edital)

Autor:

Diego Carvalho, Equipe Informática e TI, Evandro Dalla Vecchia Pereira , Pedro Henrique Chagas Freitas, Thiago Rodrigues

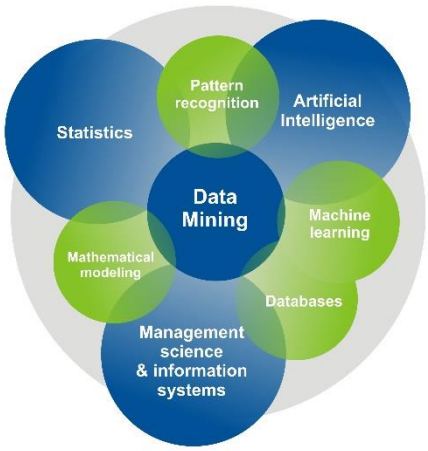
Cavalcanti, Raphael Henrique Lacerda

08 de Outubro de 2021

THIAGO CAVALCANTI
PROFESSOR

MINERAÇÃO DE DADOS

Existe geralmente informação "**escondida**" nos dados que não são tão evidentes no momento da leitura. Um analista humano pode **levar semanas** para descobrir essa informação útil. A maioria dos dados, de fato, nunca é analisada.



Com o crescimento da capacidade de **processamento e de armazenamento** surgem perguntas sobre como identificar padrões ("X" acontece se...), exceções (isto é diferente de... por causa de...), tendências (ao longo do tempo, "Y" deve acontecer...) e correlações (se "M" acontece, "N" também deve acontecer).

A mineração de dados trata da solução de problemas, analisando dados já presentes em bancos de dados. Suponha um problema de lealdade inconstante dos clientes em um mercado altamente competitivo. Em outras palavras, como verificar se seu cliente troca de fornecedor toda hora. Um banco de dados de clientes, juntamente com os perfis dos clientes, contém a chave para esse problema. Os padrões de comportamento dos antigos clientes podem ser analisados para **identificar as características das pessoas com probabilidade de mudar de produto e aquelas que provavelmente permanecerão fiéis a sua marca**.

Uma vez que essas características sejam encontradas, elas podem ser postas em prática para identificar os clientes atuais que provavelmente abandonarão o barco. Este grupo pode ser alvo de um tratamento especial, tratamento muito caro para aplicar à base de clientes como um todo.

As mesmas técnicas podem ser usadas para identificar clientes que possam ser atraídos por outro serviço que a empresa oferece ou para direcioná-los para ofertas especiais. Na economia altamente competitiva, centrada no cliente e orientada a serviços de hoje, **os dados são a matéria-prima que alimenta o crescimento dos negócios**.

A **mineração de dados** é definida como o processo de **descoberta de padrões nos dados**. O processo deve ser automático ou (mais comumente) semiautomático. Os padrões

descobertos devem ser significativos, pois levam a alguma vantagem competitiva - por exemplo, uma vantagem econômica. Perceba que os dados estão invariavelmente presentes em quantidades substanciais. A verdade é:

> Padrões úteis nos permitem fazer previsões não triviais sobre novos dados.

Há dois extremos para a expressão de um padrão: **uma caixa-preta** cujas entranhas são efetivamente incompreensíveis e **uma caixa transparente** cuja construção revela a **estrutura** do padrão. Ambos, estamos assumindo, fazem boas previsões. A diferença é se os padrões que são extraídos são ou não representados em termos de uma estrutura que pode ser examinada, fundamentada e usada para informar decisões futuras. Tais padrões nós chamamos de **estruturais** porque eles capturam a estrutura de decisão de uma maneira explícita. Em outras palavras, eles ajudam a explicar algo sobre os dados.

Nesta aula vamos tratar de tarefas e técnicas que fazem parte deste universo que tenta entender e capturar padrões sobre uma quantidade relativamente grande de dados. Muitas das técnicas que abordamos se desenvolveram em um campo conhecido como **aprendizado de máquina**. Vamos aproveitar essa introdução ao assunto para contextualizar o termo dentro do tema mineração de dados.

Aprendizado de máquina

O que é aprender, afinal? O que é aprendizado de máquina? Estas são questões filosóficas e não nos interessamos muito por filosofia nessa aula; nossa ênfase está focada na sua prova de concurso. No entanto, vale a pena passar alguns instantes tratando sobre questões fundamentais, apenas para ver o quão complicado elas são, antes de arregaçar as mangas e olhar para a aprendizagem de máquina na prática. Nosso dicionário define “aprender” como:

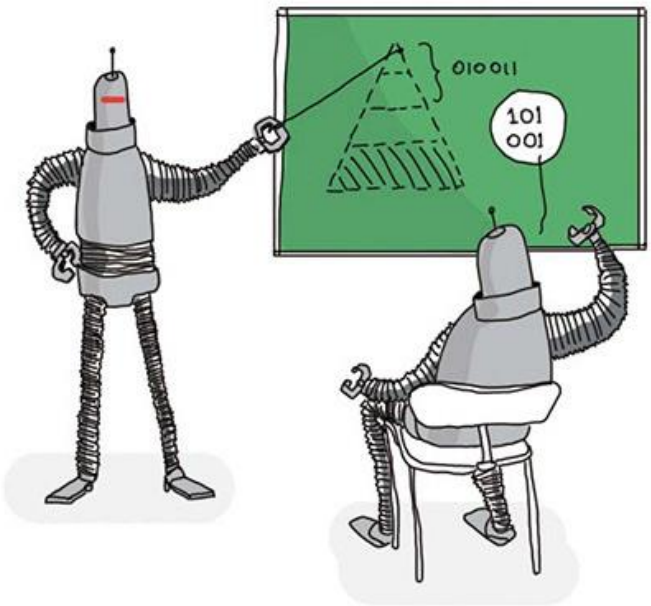
1. Ficar sabendo, reter na memória, tomar conhecimento de,
2. Adquirir habilidade prática (em),
3. Passar a compreender (algo) melhor graças a um depuramento da capacidade de apreciação, empatia, percepção etc.

Esses significados têm algumas falhas quando associados a computadores. Veja se você consegue responder a seguinte pergunta: como sabemos se uma máquina "tem conhecimento sobre " alguma coisa? Toda a questão sobre se os computadores podem estar cientes ou conscientes é uma questão filosófica. O fato é, será que eles conseguem aprender?

Anteriormente, definimos mineração de dados operacionalmente, como o processo de descoberta de padrões, de forma automática ou semiautomática, em grandes quantidades de dados – e que esses padrões devem ser úteis. Uma definição operacional pode ser formulada da mesma maneira para a aprendizagem. **As coisas são aprendidas quando eles mudam o comportamento de uma forma que nos faz ter um melhor desempenho no futuro.**

Tal fato associa o aprendizado ao **desempenho** e não ao conhecimento. Você pode testar o aprendizado observando o comportamento e comparando-o com o comportamento passado. Este é um tipo de definição muito mais objetiva e parece ser muito mais satisfatória.

Mas ainda há um problema. A aprendizagem é um conceito bastante escorregadio. Muitas coisas mudam seu comportamento de forma a torná-las melhor no futuro, mas não queremos dizer que elas realmente aprenderam. Um bom exemplo é um chinelo confortável. Será que ele aprendeu a forma do seu pé? Certamente mudou sua forma para se tornar melhor como um chinelo! No entanto, dificilmente podemos chamar isso de aprendizado.



TREINAMENTO X APRENDIZADO

Na linguagem cotidiana, muitas vezes usamos a palavra "**treinamento**" para denotar um tipo de aprendizado sem sentido. Nós treinamos animais e até plantas. Mas aprender é diferente. **Aprender implica pensar**. Aprender implica **propósito**. Algo que se aprende tem que ser feito intencionalmente. É por isso que não falamos que uma vinha aprendeu a crescer em torno de uma treliça em um vinhedo - falamos que ela foi treinada. Aprender sem propósito é apenas treinar.

Assim, em um exame mais detalhado, uma definição de aprendizado, em termos operacionais e orientados para o desempenho, tem seus próprios problemas quando falamos sobre computadores. Para decidir se algo realmente aprendeu, você precisa ver o

que se pretendia, se havia algum propósito envolvido. Isso torna o conceito discutível quando aplicado a máquinas porque não é claro se os artefatos se comportam propositadamente. Enfim ... discussões filosóficas sobre o que realmente significa "aprender", como discussões sobre o que realmente significa "intenção" ou "propósito" estão repletas de dificuldades. Até os tribunais de justiça acham difícil lidar com a intenção.

Felizmente, os tipos de técnicas de aprendizado explicadas nesta aula não apresentam esses problemas conceituais - eles são chamados de "**aprendizado de máquina**" sem realmente pressupor qualquer posição filosófica específica sobre o que a aprendizagem realmente é. A **mineração de dados** é um tópico prático e envolve aprendizado em um sentido prático, não teórico. Estamos interessados **em técnicas para encontrar padrões em dados, padrões que forneçam insight ou possibilitem tomadas de decisão rápidas e precisas**.

Os dados tomarão a forma de um conjunto de exemplos - exemplos de clientes que trocaram a lealdade ou situações em que certos tipos de lentes de contato podem ser prescritos. A saída toma a forma de previsões sobre novos exemplos - uma previsão sobre se um determinado cliente mudará ou uma previsão de que tipo de lente será prescrita sob determinadas circunstâncias.

Muitas técnicas de aprendizado procuram descrições estruturais do que é aprendido, descrições que podem se tornar bastante complexas e são tipicamente expressas como **conjuntos de regras ou como árvores de decisão**. Como elas podem ser entendidas pelas pessoas, essas descrições servem para explicar o que foi aprendido, em outras palavras, para **explicar a base para novas previsões**.

A experiência mostra que, em muitas aplicações de aprendizado de máquina para mineração de dados, as estruturas de conhecimento explícitas que são adquiridas, essas

descrições estruturais, são pelo menos tão importantes quanto a capacidade de ter um bom desempenho em novos exemplos. As pessoas frequentemente usam a mineração de dados para obter conhecimento, não apenas previsões. Obter conhecimento a partir de dados certamente parece uma boa ideia se você puder fazê-lo.

Para ajudar a resolver esses questionamentos surge um conjunto de conceitos relacionados à Data Mining e aprendizado de máquina. Veremos a estrutura teórica da matéria nesta parte da aula. Vem comigo! A figura abaixo apresenta alguns conceitos sobre aprendizados que serão vistos mais à frente, não tente entendê-los agora. Deixe apenas a sua mente capturar uma primeira percepção sobre o assunto.

ESQUEMATIZANDO

CONCEITOS BÁSICOS: MINERAÇÃO DE DADOS

Começamos os conceitos tentando responder ao seguinte questionamento: O que é mineração de dados? Vários autores propuseram definições semelhantes para o termo, vejamos algumas delas:

1. Eduardo Gimenes: É o processo de **extrair informação válida**, previamente desconhecida e de máxima abrangência a partir de **grandes bases de dados**, usando-as para efetuar **decisões** cruciais.
2. Laudon&Laudon: **Análise** de grandes quantidades de **dados** a fim de encontrar **padrões e regras** que possam ser usadas para orientar a **tomada de decisões** e **prever o comportamento** futuro.
3. Mineração de dados, ou data mining, é o processo de análise de conjuntos de dados que tem por objetivo a descoberta de padrões interessantes e que possam representar informações úteis.

Como o nome indica, data mining se refere à mineração ou à descoberta de novas informações em função de padrões ou regras em grandes quantidades de dados. Para ser útil, na prática, a mineração de dados precisa ser realizada eficientemente em grandes arquivos e banco de dados.

Vejamos outras definições possíveis para o termo. “A mineração de dados é a aplicação de algoritmos específicos para extração de padrões a partir dos dados” de FAYYAD. “A

mineração de dados se refere à extração, ou mineração, de conhecimento a partir de grandes quantidades de dados" de HAN e KAMBER.

As definições não param por aí ... nós vamos ver abaixo diversas definições para Data Mining que já caíram em prova – é importantíssimo saber esse conceito, visto que as bancas adoram cobrar as maneiras diferentes de se definir mineração de dados. Vamos lá ...

DEFINIÇÕES

Data Mining é o processo de explorar grande quantidade de dados para extração não-trivial de informação implícita desconhecida.

Palavras-chave: exploração; informação implícita desconhecida.

Data Mining é uso de teorias, métodos, processos e tecnologias para organizar uma grande quantidade de dados brutos para identificar padrões de comportamentos em determinados públicos.

Palavras-chave: teorias; métodos; processos; tecnologias; organizar dados brutos; padrões de comportamentos.

Data Mining é a categoria de ferramentas de análise denominada open-end e que permite ao usuário avaliar tendências e padrões não conhecidos entre os dados.

Palavras-chave: ferramenta de análise; open-end; tendências e padrões.

Data Mining é o processo de descoberta de novas correlações, padrões e tendências entre as informações de uma empresa, por meio da análise de grandes quantidades de dados armazenados em bancos de dados usando técnicas de reconhecimento de padrões, estatísticas e matemáticas.

Palavras-chave: descoberta; correlações; padrões; tendências; reconhecimento de padrões; estatística; matemática.

Data Mining constitui em uma técnica para a exploração e análise de dados, visando descobrir padrões e regras, a princípio ocultos, importantes à aplicação.

Palavras-chave: exploração e análise de dados; padrões; regras; ocultos.

Data Mining é o conjunto de ferramentas que permitem ao usuário avaliar tendências e padrões não conhecidos entre os dados. Esses tipos de ferramentas podem utilizar técnicas avançadas de computação como redes neurais, algoritmos genéticos e lógica nebulosa (fuzzy), dentre outras.

Palavras-chave: tendências; padrões; redes neurais; algoritmos genéticos; lógica nebulosa.

Data Mining é o conjunto de ferramentas e técnicas de mineração de dados que têm por objetivo buscar a classificação e o agrupamento (clusterização) de dados, bem como identificar padrões.

Palavras-chave: classificação; agrupamento; clusterização; padrões.

Data Mining é o processo de explorar grandes quantidades de dados à procura de padrões consistentes com o intuito de detectar relacionamentos sistemáticos entre variáveis e novos subconjuntos de dados.

Palavras-chave: padrões; relacionamentos.

Data Mining consiste em explorar um conjunto de dados visando a extrair ou a ajudar a evidenciar padrões, como regras de associação ou sequências temporais, para detectar relacionamentos entre estes.

Palavras-chave: exploração; padrões; regras; associação; sequência temporal; detecção.

Data Mining são ferramentas que utilizam diversas técnicas de natureza estatística, como a análise de conglomerados (cluster analysis), que tem como objetivo agrupar, em diferentes conjuntos de dados, os elementos identificados como semelhantes entre si, com base nas características analisadas.

Palavras-chave: estatística; análise de conglomerados; agrupamento.

Data Mining é o conjunto de técnicas que, envolvendo métodos matemáticos e estatísticos, algoritmos e princípios de inteligência artificial, tem o objetivo de descobrir relacionamentos significativos entre dados armazenados em repositórios de grandes volumes e concluir sobre padrões de comportamento de clientes de uma organização.

Palavras-chave: métodos matemáticos e estatístico; inteligência artificial; relacionamentos; padrões; comportamentos.

Data Mining é o processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados.

Palavras-chave: padrões; regras de associação; sequências temporais; relacionamentos.

Data Mining é o processo de identificar, em dados, padrões válidos, novos, potencialmente úteis e, ao final, compreensíveis.

Palavras-chave: padrões; utilidade.

Data Mining é um método computacional que permite extrair informações a partir de grande quantidade de dados.

Palavras-chave: extração.

Data Mining é o processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais.

Palavras-chave: exploração; padrões consistentes; regras de associação; sequência temporal.

Data Mining é o processo de analisar de maneira semiautomática grandes bancos de dados para encontrar padrões úteis.

Palavras-chave: padrões.

Todas as definições acima caíram em prova – sem exceção. O professor Diego Carvalho fez essa listagem baseado em provas anteriores. **Notem como é importante saber de forma abrangente as possíveis definições para mineração de dados**. Dito isso, vamos tentar condensar todas elas em uma grande definição a seguir:

> *Data Mining – Mineração de Dados – é um conjunto de processos, métodos, teorias, ferramentas e tecnologias open-end utilizadas para explorar, organizar e analisar de forma semiautomática uma grande quantidade de dados brutos com o intuito de identificar, descobrir, extrair, classificar e agrupar informações implícitas desconhecidas, além de avaliar correlações, tendências e padrões consistentes de comportamento potencialmente úteis – como regras de associação ou sequências temporais – de forma não-trivial por meio de técnicas estatísticas e matemáticas, como redes neurais, algoritmos genéticos, inteligência artificial, lógica nebulosa, análise de conglomerados (clusters), entre outros.*

Com o crescente avanço da quantidade de dados em todas as aplicações, a mineração de dados atende à necessidade iminente de uma análise de dados eficaz, escalável e flexível em nossa sociedade. A mineração de dados pode ser considerada como uma evolução natural da tecnologia da informação e uma confluência de várias disciplinas relacionadas e domínios de aplicação.

A mineração de dados é, portanto, o processo de descobrir padrões interessantes a partir de grandes quantidades de dados. Como um **processo** de descoberta de conhecimento, normalmente envolve **limpeza de dados, integração de dados, seleção de dados, transformação de dados, descoberta de padrões, avaliação de padrões e apresentação de conhecimento**.

Segundo HAN, uma visão multidimensional da mineração de dados descreve as principais dimensões como sendo dados, conhecimento, tecnologias e aplicativos.

A mineração de dados pode ser realizada em **qualquer tipo de dados**, desde que os dados sejam significativos para um aplicativo de destino, como dados do banco de dados, dados do data warehouse, dados transacionais e tipos de dados avançados. Os tipos de dados avançados incluem dados relacionados a tempo ou sequência, fluxos de dados, dados espaciais e espaço-temporais, dados de texto e multimídia, dados em gráfico e em rede e dados da Web.

Um **data warehouse** é um repositório para armazenamento em longo prazo de dados de várias origens, organizado de modo a facilitar a tomada de decisões gerenciais. Os dados são armazenados em um esquema unificado e geralmente são resumidos ou agregados. Os sistemas de data warehouse fornecem recursos de análise de dados multidimensionais, coletivamente chamados de processamento analítico on-line (OLAP).

A mineração de dados multidimensional (também chamada de mineração de dados multidimensional exploratória) integra as técnicas de mineração de dados com a análise multidimensional baseada em OLAP. Ela procura padrões interessantes entre múltiplas combinações de dimensões (atributos) em níveis variados de abstração, explorando, assim, o espaço de dados multidimensional.

A mineração de dados, como um domínio altamente orientado a aplicativos, incorporou tecnologias de muitos outros domínios. Estes incluem **estatísticas**, **aprendizado de máquina**, **bancos de dados** e **sistemas de armazenamento de dados e recuperação de informações**. A natureza interdisciplinar da pesquisa e desenvolvimento de mineração de dados contribui significativamente para o sucesso da mineração de dados e suas extensas aplicações.

Antes de continuarmos, vamos então fazer uma questão CESPE sobre esses conceitos básicos para fixarmos o assunto.

HORA DE PRATICAR!

(Ano: 2016 Banca: CESPE Órgão: TCE-PA Prova: Auditor de Controle Externo - Área Informática - Analista de Sistemas) Julgue o item a seguir, em relação a data warehouse e data mining.

No contexto de data mining, o processo de descoberta de conhecimento em base de dados consiste na extração não trivial de conhecimento previamente desconhecido e potencialmente útil.

_____

Comentário: Após observar uma lista de definições do conceito de data-mining, podemos afirmar que essa alternativa está correta.

Gabarito: CERTO.

(Ministério da Economia – Especialista em Ciência de Dados - 2020) No que se refere à mineração de dados, julgue os itens a seguir.

Mecanismos de busca utilizam mineração de textos para apresentar ao usuário os resultados de suas pesquisas, de modo que ambos os conceitos se equivalem.

_____

Comentários: A tecnologia de mineração de textos não é um mecanismo de busca, pois a mineração ajuda o usuário a descobrir informações previamente desconhecidas, enquanto na busca o usuário já sabe o que deseja procurar.

Gabarito: ERRADO.

A mineração de dados tem muitas aplicações de sucesso, como na inteligência de negócios, pesquisa na Web, bioinformática, informática em saúde, finanças, bibliotecas digitais e governos digitais. Existem, ainda, muitos problemas desafiadores na pesquisa de mineração de dados. As áreas incluem metodologia de mineração, interação com o usuário, eficiência e escalabilidade, além de lidar com diversos tipos de dados. A pesquisa de mineração de dados tem impactado fortemente a sociedade e continuará a fazê-lo no futuro.

Problemas onde as técnicas tradicionais poderiam não se ajustar à enorme quantidade de dados, à alta dimensionalidade dos dados e à heterogeneidade e natureza distribuídas dos dados podem aparecer no radar. Data Mining, então, surge para completar essa lacuna. Sua base teórica é uma mistura de diferentes disciplinas, como já falamos.

Ok! Então, antes de começarmos a descrever o que mais faz parte da mineração de dados, vamos tratar do que alguns autores conhecem como falácias de Data Mining. São basicamente quatros:

Falácias de Data Mining

Data Mining é automático: é um processo, é iterativo, requer supervisão. (observação: embora em algumas definições que vimos no início da aula apareça o termo automático, não há consenso entre os autores quanto ao aspecto automático do processo de mineração.)

Investimentos são recuperados rapidamente: depende de muitos fatores!

Software são intuitivos e simples: é mais importante conhecer os conceitos dos algoritmos e o negócio em si!

Data Mining pode identificar problemas no negócio: DM pode encontrar padrões e fenômenos, identificar a causa deve ser feito por especialistas.

Podemos concluir, então, que existe um processo iterativo, que requer supervisão e depende de vários fatores para uma implementação de sucesso. É importante entender os algoritmos, as tarefas e o negócio. Assim, é possível encontrar os padrões e fenômenos sobre a massa de dados.

O uso da mineração de dados é, portanto, potencializada por alguns fatores: o volume de dados disponível atualmente é enorme, o fato de os dados estarem mais organizados, os recursos computacionais estão cada vez mais potentes, a competição empresarial exige técnicas mais modernas de decisão e os programas comerciais de mineração de dados já podem ser adquiridos.

Para executarmos qualquer análise sobre os dados é necessário que tenhamos em mente qual a tarefa que estamos realizando. Uma **tarefa de mineração de dados** consiste na especificação **do que** queremos buscar nos dados. Podemos buscar por algum tipo de regularidade ou categoria de padrões que temos interesse em encontrar ou ainda padrões que poderiam nos surpreender (por exemplo, um gasto exagerado de um cliente de cartão de crédito, fora dos padrões usuais de seus gastos).

A classificação das tarefas pode ser feita de acordo com alguns critérios. O primeiro divide as tarefas em **descritivas** e **preditivas**. As **descritivas** caracterizam as propriedades gerais dos dados em um banco de dados. Estão focadas em achar padrões reconhecidos por seres humanos para descrever os dados. As **preditivas**, por outro lado, realizam uma inferência sobre os dados atuais para fazer previsões futuras sobre eles. Usa variáveis para prever valores futuros ou desconhecidos de outras variáveis.

(Ministério da Economia – Especialista em Ciência de Dados - 2020) No que se refere à mineração de dados, julgue os itens a seguir.

Modelagem preditiva é utilizada para antecipar comportamentos futuros, por meio do estudo da relação entre duas ou mais variáveis

_____

Comentários: A modelagem é preditiva quanto tentamos estimar valores futuros. Neste tipo de modelagem não temos certeza sobre o resultado obtido pelo modelo e a medição dos dados (no futuro). Isso nos leva a uma probabilidade de estarmos ou não certos em relação a nossa previsão. Assim, temos uma alternativa correta.

Gabarito: CERTO.

Outra taxonomia divide as tarefas em top-down e botton-up. Algumas tarefas são abordadas de forma **top-down** chamado **teste de hipóteses**. Em testes de hipóteses, um comportamento armazenado no banco de dados passado é utilizado para verificar ou refutar notações preconcebidas, ideias e palpites referentes às relações nos dados.

Outras tarefas são melhor abordadas de forma bottom-up chamado de descoberta de conhecimento (Knowledge discorvery). Na descoberta de conhecimento, a análise sobre os dados é feita sem suposições prévias. Os dados são autorizados a falar por si.

As tarefas adequadas para mineração de dados (não é limitado a essas) são:

Classificação (Preditiva)
Clustering (Descritiva)
Regra de Associação (Descritiva)
Regressão (Preditiva)
Detecção de desvios (Preditiva).

Outro ponto importante dentro do assunto são as técnicas de mineração que consistem na especificação de **métodos** que nos garantam como descobrir os padrões que nos interessam. Dentre as principais técnicas utilizadas em mineração de dados, temos: técnicas estatísticas, técnicas de aprendizado de máquina e técnicas baseadas em crescimento-poda-validação.

Por fim, temos três características que são aplicadas a muitos conjuntos de dados e que possuem um impacto significativo sobre as técnicas de mineração de dados: **dimensão**, **dispersão** e **resolução**. A **dimensão** refere-se à quantidade de atributos de um conjunto de dados. A **resolução** está relacionada à granularidade dos dados. Um conjunto de dados é muito disperso quando para um atributo relevante, a maioria dos valores é NULL (desconhecido) ou um valor padrão. Esse conceito está relacionado à **dispersão.**

Um último conceito que já foi cobrado em provas de concurso diz respeito aos métodos para identificar padrões em dados, que são basicamente três:

- **Modelos simples** (consultas baseadas em SQL, OLAP, raciocínio humano)
- **Modelos intermediários** (regressão, árvores de decisão, agrupamento)
- **Modelos complexos** (redes neurais, indução de regras)

OBJETIVOS

Segundo Navathe, a mineração de dados costuma ser executada com alguns objetivos finais ou aplicações. De um modo geral, esses objetivos se encontram nas seguintes classes: Previsão, Identificação, Classificação ou Otimização. *Como assim?* Cara, isso significa que você pode utilizar a mineração de dados com o objetivo de prever, identificar, classificar ou otimizar algo. *Tranquilo?* Fiquem com o mnemônico abaixo para ajudar a memorizar:

Previsão

A mineração de dados pode mostrar como certos atributos dos dados se comportarão no futuro. Um de seus objetivos é prever comportamentos futuros baseado em comportamentos passados. Exemplo: análise de transações de compras passadas para prever o que os consumidores comprarão futuramente sob certos descontos, quanto volume de vendas uma loja gerará em determinado período e se a exclusão de uma linha de produtos gerará mais lucros.

Identificação

Padrões de dados podem ser usados para identificar a existência de um item, um evento ou uma atividade. Por exemplo: intrusos tentando quebrar um sistema podem ser identificados pelos programas por eles executados, arquivos por eles acessados ou pelo tempo de CPU por sessão aberta. Em aplicações biológicas, a existência de um gene pode ser identificada por sequências específicas de nucleotídeos em uma cadeia de DNA.

Classificação

A mineração de dados pode particionar os dados de modo que diferentes classes ou categorias possam ser identificadas com base em combinações de parâmetros. Por exemplo: os clientes em um supermercado podem ser categorizados em compradores que buscam desconto, compradores com pressa, compradores regulares leais, compradores ligados a marcas conhecidas e compradores eventuais.

Otimização

Um objetivo relevante da mineração de dados pode ser otimizar o uso de recursos limitados, como tempo, espaço, dinheiro ou materiais e maximizar variáveis de saída como vendas ou lucros sob determinado conjunto de restrições. Exemplo: a execução de um projeto que deve respeitar completamente o orçamento contratado, o escopo combinado e o cronograma previsto de forma a maximizar o resultado.

(CESPE – TCE/SC – Auditor Fiscal de Controle Externo) Para a realização de prognósticos por meio de técnicas de mineração de dados, parte-se de uma série de valores existentes obtidos de dados históricos bem como de suposições controladas a respeito das condições futuras, para prever outros valores e situações que ocorrerão e, assim, planejar e preparar as ações organizacionais.

_____
Comentários: conforme vimos em aula, prognóstico ou previsão partem dados históricos para prever situações futuras (Correto).

(ESAF – ANAC – Analista Administrativo) São objetivos da Mineração de Dados:

a) Distribuição, Identificação, Organização e Otimização.
b) Previsão, Priorização, Classificação e Alocação.
c) Previsão, Identificação, Classificação e Otimização.
d) Mapeamento, Identificação, Classificação e Atribuição.
e) Planejamento, Redirecionamento, Classificação e Otimização.

_____
Comentários: conforme vimos em aula, é a Previsão, Identificação, Classificação e Otimização (Letra C).

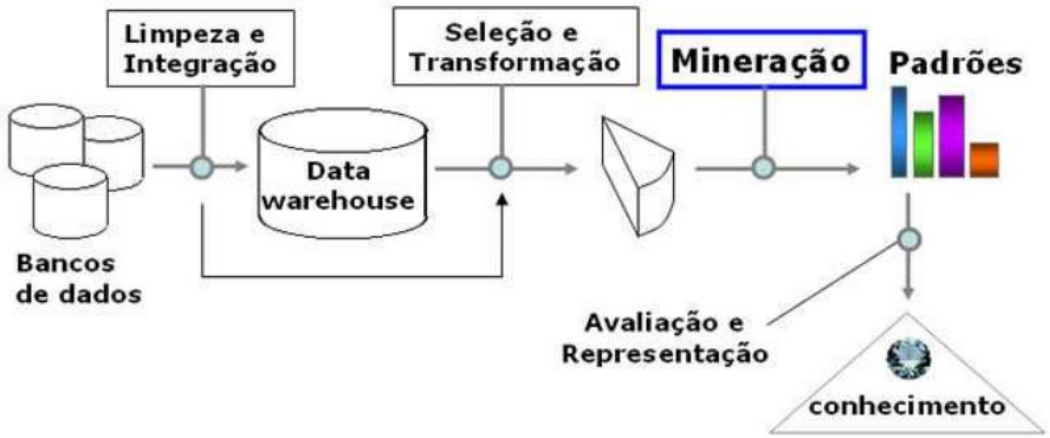
(FUNCAB – MDA – Administrador de Banco de Dados) A mineração de dados costuma ser executada com alguns objetivos finais ou aplicações. Em geral, esses objetivos se encontram nas seguintes classes:

a) levantamento, previsão, classificação e otimização.
b) requisito, identificação, classificação e otimização.
c) previsão, identificação, levantamento e requisito
d) levantamento, requisito, classificação e otimização.
e) previsão, identificação, classificação e otimização

_____
Comentários: conforme vimos em aula, é a Previsão, Identificação, Classificação e Otimização (Letra E).

PROCESSO DE MINERAÇÃO

Antes de falar do processo de mineração propriamente dito, vamos examinar a figura abaixo que trata do processo de BI.



Baseado na figura é possível listar as etapas do processo:

1. Limpeza dos dados: etapa onde são eliminados ruídos e dados inconsistentes.

2. Integração dos dados: etapa onde diferentes fontes de dados podem ser combinadas, produzindo um único repositório de dados.

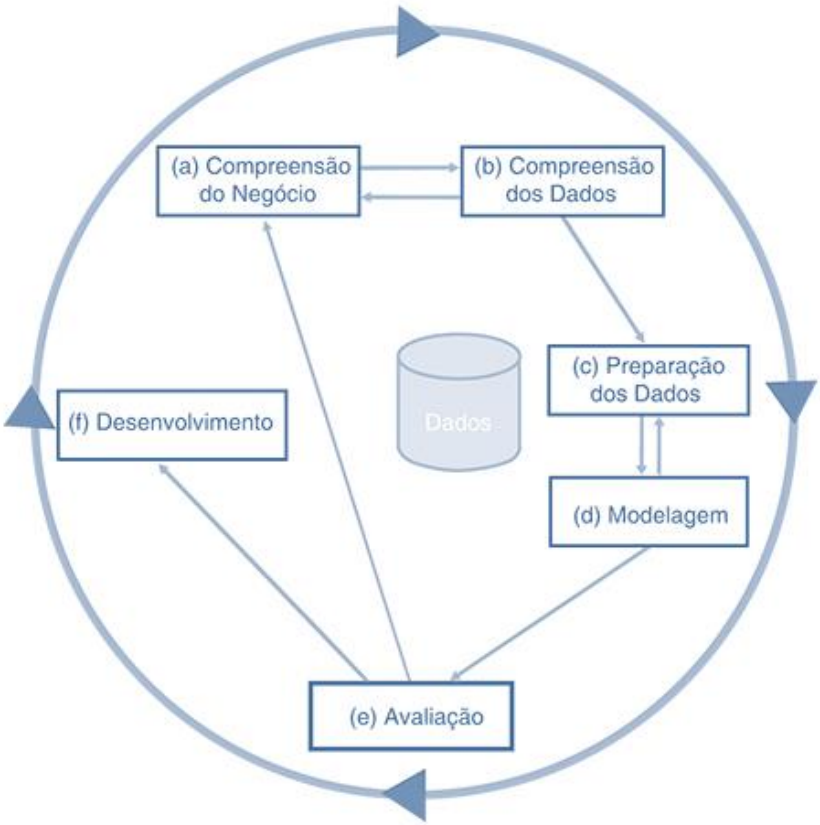
3. Seleção: etapa onde são selecionados os atributos que interessam ao usuário. Por exemplo, o usuário pode decidir que informações como endereço e telefone não são relevantes para decidir se um cliente é um bom comprador ou não.

4. Transformação dos dados: etapa onde os dados são transformados num formato apropriado para aplicação de algoritmos de mineração (por exemplo, por meio de operações de agregação).

5. Mineração: etapa essencial do processo consistindo na aplicação de técnicas inteligentes a fim de se extrair os padrões de interesse.

CRISP-DM

O processo de mineração de dados se assemelha um pouco ao processo de BI descrito acima. Propõe uma **visão geral do ciclo de vida** de um **projeto** de mineração de dados. Vejam a figura abaixo:

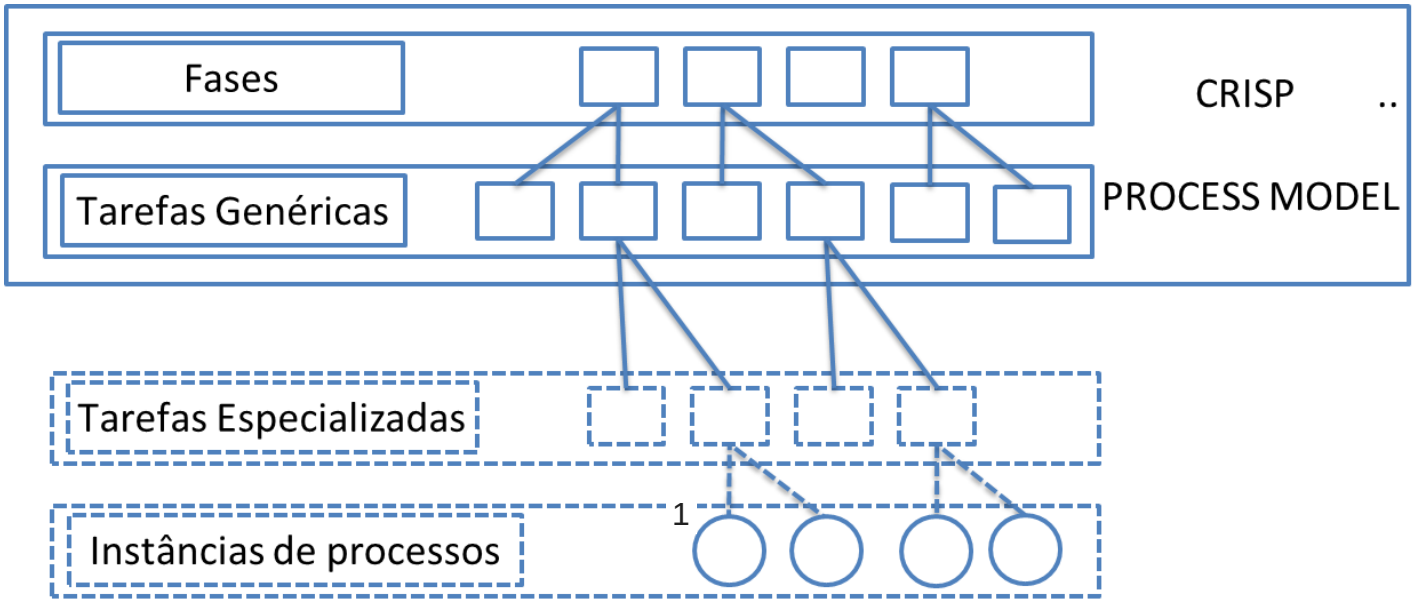


Em 1996, um conjunto de três empresas especializadas no então jovem e imaturo mercado de data mining desenvolveram um modelo de processos genéricos, com o intuito de padronizar as etapas do processo de mineração de dados, dando início ao denominado projeto CRISP-DM (CRoss Industry Standard Process for Data Mining) [The CRISP-DM Consortium, 2000].

Na figura acima é mostrado o ciclo de vida de um projeto de mineração de dados, que consiste em seis fases. A sequência de fases não é obrigatória, ocorrendo a transição para diferentes fases, dependendo do resultado de cada fase, e qual etapa particular de cada fase precisa ser executada em seguida. As setas indicam as mais importantes e mais frequentes dependências entre as fases.

O ciclo externo na figura simboliza o ciclo natural da mineração de dados. Um processo de mineração de dados continua após a solução ter sido desenvolvida. As lições aprendidas durante o processo podem provocar perguntas novas, frequentemente mais pertinentes ao negócio. Processos subsequentes se beneficiarão das experiências de processos anteriores.

Repartição em Quatro Níveis da Metodologia CRISP-DM



Na figura acima temos hierarquia de estruturação da metodologia CRIP-DM. No nível superior, o processo de mineração de dados é organizado em várias fases. **Cada fase consiste em várias tarefas genéricas de segundo nível**. Esse segundo nível é chamado genérico porque pretende ser geral o suficiente para cobrir **todas as situações possíveis de mineração de dados**.

As tarefas genéricas devem ser o **mais completas e estáveis possível**. Por completos queremos dizer que abrangem todo o processo de mineração de dados e todos os aplicativos possíveis de mineração de dados. Já estável significa que o modelo deve ser válido para desenvolvimentos de modelos imprevisíveis, utilizando novas técnicas de modelagem.

O terceiro nível, o nível de tarefa especializada, é o local para **descrever como as ações das tarefas genéricas devem ser realizadas em determinadas situações específicas**. Por exemplo, no segundo nível, pode haver uma tarefa genérica chamada limpeza dos dados. O terceiro nível descreveria como essa tarefa é executada em diferentes situações, como limpeza de valores numéricos, de valores categóricos.

A descrição de fases e tarefas como etapas discretas executadas em uma ordem específica representa uma sequência idealizada de eventos. Na prática, muitas das tarefas podem ser executadas em uma ordem diferente e, muitas vezes, será necessário voltar repetidamente às tarefas anteriores e repetir determinadas ações. O modelo de processo não tenta capturar todas essas rotas possíveis por meio do processo de mineração de dados, porque isso exigiria um modelo de processo excessivamente complexo.

O quarto nível, a instância do processo, é um **registro das ações, decisões e resultados de um compromisso real de mineração de dados**. Uma instância do processo é organizada de acordo com as tarefas definidas nos níveis mais altos, mas representa o que realmente aconteceu em um determinado engajamento, e não o que acontece em geral.

Entendimento do Negócio

O entendimento do negócio (Business Understanding) foca na **compreensão do negócio** que visa obter conhecimento sobre os objetivos do negócio e seus requisitos.

Entre as principais considerações técnicas que devem ser verificadas no início de um processo de KDD (Knowledge Discovery of Database) estão:

- **Identificar as pessoas e áreas envolvidas no processo de KDD**. Os especialistas do domínio da aplicação, a equipe de tecnologia da informação e os grupos de decisão da empresa devem ser submetidos, sempre que necessário, a um treinamento em KDD que nivele o conhecimento técnico na área.

- **Esboçar uma lista de necessidades e expectativas dos indivíduos envolvidos** quanto ao propósito do 6 DD. Para isso, é necessário **identificar que critérios** podem ser adotados para mensurar o sucesso do processo.

- **Fazer um levantamento do hardware e software existentes.** O CRISP-DM recomenda que o processo de KDD seja realizado em plataforma com arquitetura expansível, que suporte grandes volumes de dados, boa capacidade de processamento e acesso à base de dados heterogênea.

- **Fazer um inventário das bases de dados disponíveis**. Incluem-se neste item tanto bases de dados internas quanto externas. Potenciais bases externas são aquelas relacionadas com o domínio da aplicação e que possam ser utilizadas no enriquecimento dos dados.

- **Verificar a existência de Data Warehouses**. A utilização de Data Warehouses preexistentes pode poupar muitos esforços de pré-processamento dos dados.

- **Identificar e documentar, se possível, todo o conhecimento prévio** existente e disponível acerca do **domínio da aplicação**. Este conhecimento pode ser útil no decorrer do processo.

Uma vez definido o domínio sobre o qual se pretende executar o processo de descoberta, o próximo passo é selecionar e coletar o conjunto de dados ou variáveis necessárias. Consiste no entendimento dos dados utilizando-se de conjuntos de dados "modelo". Antes de continuarmos vajamos uma rápida questão sobre o assunto:

(Ministério da Economia – Especialista em Ciência de Dados - 2020) No que se refere à mineração de dados, julgue os itens a seguir.

No modelo CRISP-DM, a fase na qual se planejam todas as atividades para carga dos dados é denominada entendimento dos dados.

Comentários: O entendimento das atividades é feito na fase **de entendimento dos negócios**.

Gabarito: ERRADO.

Compreensão dos Dados

Essa fase se inicia com **uma coleta inicial de dados**, e com procedimentos e atividades visando **à familiarização com os dados**, para identificar possíveis problemas de qualidade, ou detectar subconjuntos interessantes para formar hipóteses. É geralmente executada em conjunto com a fase anterior, e requer um **estudo minucioso das informações disponíveis**. Esta fase normalmente envolve as seguintes atividades:

- **Compreender o significado e perceber a relevância dos atributos/ dados disponíveis**. Esta análise tem um papel crucial na definição dos objetivos do processo de KDD. Metadados acerca das bases de dados e seus atributos devem ser documentados.

- **Avaliar a qualidade dos dados disponíveis**. Deve-se procurar identificar o propósito para o qual os dados foram coletados, assim como a caracterização do nível de ruído envolvido. Bases de dados poluídas requerem o uso de ferramentas adequadas ao processo de limpeza. Em geral, os recursos de limpeza são fortemente dependentes do domínio da aplicação e podem ser otimizados com a utilização de conhecimentos preexistentes.

- **Verificar se os dados estão disponíveis em quantidade suficiente para o processo de KDD**. Bases de dados pequenas ou que contenham dados pouco representativos das condições normais do domínio da aplicação podem inviabilizar o processo de KDD.

A próxima etapa é a preparação dos dados (Data Preparation) que consiste na **limpeza**, **transformação**, **integração** e **formatação** dos dados da etapa anterior. É a atividade pela qual os ruídos, dados estranhos ou inconsistentes são tratados. Esta fase abrange todas as atividades para construir o conjunto de dados final (dados que serão alimentados nas ferramentas de mineração), a partir do conjunto de dados inicial.

Preparação dos dados

A Preparação dos Dados compreende as ações de pré-processamento dos dados para a fase de modelagem propriamente dita. São exemplos de ações desta fase:

- **Selecionar os dados que serão efetivamente analisados**. A seleção de dados pode integrar dados coletados de diferentes fontes, enriquecendo o conjunto de dados que será analisado.

- **Promover a limpeza dos dados**, procurando remover inconsistências e completar (ou eliminar) dados ausentes.

- **Adequar o formato** dos dados.
- **Construir novos atributos** a partir de atributos existentes.

Lembre-se: A utilização de Data Warehouses facilita esta etapa do processo de mineração de dados, que costuma ser a fase que exige mais esforço, correspondendo geralmente a mais de 50% do trabalho. Por isso, é muito importante para uma organização que ela possua em seus processos habituais boas práticas da administração de dados, como o *Data Cleansing*, que é uma parte fundamental da cadeia da admir...tração da informação, responsável pelas etapas de detecção, validação e correção de erros em bases de dados.

Vejamos uma questão sobre o assunto:

(Ano: 2015 Banca: CESPE Órgão: TJ-DFT Prova: Técnico Judiciário - Programação de Sistemas) Julgue o item a seguir, a respeito de data warehouse e de data mining.

Em um processo de mineração, durante a etapa de preparação dos dados, são analisados os requisitos de negócio para consolidar os dados.

Comentário: Vejam que o processo tem uma etapa específica para analisar os requisitos de negócio. Em outra etapa, temos **a preparação dos dados.** Logo, a afirmação acima está incorreta.

Gabarito: E.

Modelagem dos Dados

Esta fase consiste na escolha e na aplicação da(s) técnica(s) de modelagem (algoritmo(s) de mineração) sobre os dados a serem analisados. Corresponde à etapa de Mineração de Dados do Processo de KDD. Envolve testes iniciais voltados à calibração de parâmetros do(s) algoritmo(s).

Diversas técnicas de modelagem dos dados são experimentadas e, em cada uma delas, diversos valores de parâmetros são testados. Essa atividade prevê um retorno à atividade de preparação dos dados, visto que algumas técnicas de modelagem apresentam demandas diferentes quanto ao formato do conjunto de dados utilizado.

Algumas técnicas possuem requerimentos específicos na forma dos dados. Consequentemente, voltar para a etapa de preparação de dados é frequentemente necessário. A maioria das técnicas de mineração de dados é baseada em conceitos de aprendizagem de máquina, reconhecimento de padrões, estatística.

Vejamos como esse assunto já foi cobrado em provas anteriores.

(Ministério da Economia – Especialista em Ciência de Dados - 2020) No que se refere à mineração de dados, julgue os itens a seguir.

Na etapa de mineração do data mining, ocorre a seleção dos conjuntos de dados que serão utilizados no processo de mining

Comentários: No modelo CRISP-DM, a seleção dos dados que serão efetivamente utilizados pela mineração ocorre na tarefa de **preparação dos dados.**

f

Gabarito: ERRADO.

(Ministério da Economia – Especialista em Ciência de Dados - 2020) Julgue os seguintes itens, a respeito de big data

A etapa de modelagem do modelo CRISP-DM permite a aplicação de diversas técnicas de mineração sobre os dados selecionados, conforme os formatos dos próprios dados

Comentários: A fase de modelagem de dados do CRISP-DW consiste na escolha e na aplicação da(s) técnica(s) de modelagem (algoritmo(s) de mineração) sobre os dados a serem analisados.

Gabarito: CERTO.

Avaliação do processo (Evaluation).

Neste momento, o objetivo é garantir que o modelo gerado atenda às expectativas da organização. Os resultados do processo de descoberta do conhecimento podem ser mostrados de diversas formas. Porém, estas formas devem possibilitar uma análise criteriosa para identificar a necessidade de retornar a qualquer um dos estágios anteriores do processo de mineração. Nesta etapa, verificamos se o modelo construído possui qualidade, sob uma perspectiva da análise de dados.

O resultado da fase anterior é um conjunto de um ou mais modelos de conhecimento. A fase de Avaliação compreende a verificação da qualidade desses modelos gerados diante das expectativas formuladas na fase inicial do processo. A partir dessa avaliação, os especialistas em KDD podem propor revisões nas fases anteriores e redefinir os passos seguintes a serem realizados.

Assim, antes de prosseguir, é importante avaliar mais detalhadamente o modelo, e rever as etapas executadas para construir o modelo, para se certificar de que ele conseguirá

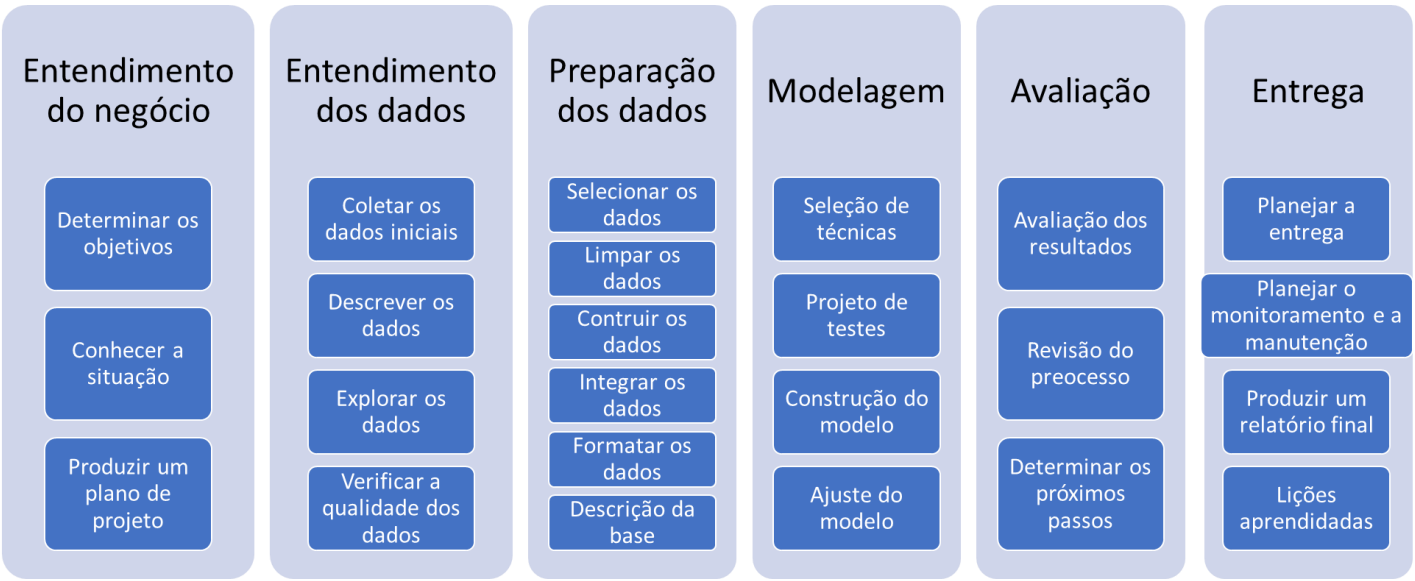
alcançar os objetivos de negócio. Deve se determinar se houve algum importante objetivo do negócio que não foi suficientemente alcançado. No fim desta fase, uma decisão sobre o uso dos resultados da mineração deve ser tomada.

Entrega (Deployment)

A Entrega (Deployment) consiste na definição das fases de implantação do projeto de Mineração de Dados. A criação do modelo não é o fim do projeto. Mesmo se a finalidade do modelo for apenas aumentar o conhecimento dos dados, o conhecimento ganho deve ser organizado e apresentado em uma maneira que o cliente possa usar.

A fase em questão consiste do planejamento e do acompanhamento das ações a serem realizadas com o(s) modelo(s) de conhecimento gerado(s) pelas fases anteriores. Essa fase envolve também a elaboração de um **relatório final do processo**, que apresenta os **resultados obtidos e possíveis alternativas** de ação de descoberta de conhecimento na organização na qual o processo esteja sendo aplicado.

A figura abaixo descreve as ações feitas em cada uma das etapas do processamento:



Dependendo das exigências, a fase de entrega pode ser tão simples quanto a geração de um relatório, ou tão complexo quanto executar processos de mineração de dados repetidamente. Em muitos casos será o cliente, não o analista dos dados, que realizará as etapas da execução. Entretanto, mesmo se o analista não se encarregar da execução é importante que ele faça o cliente compreender que medidas deverão ser tomadas a fim de empregar efetivamente os modelos criados.

Com isso, terminamos nossa rápida explicação sobre o processo de mineração de dados descrito pelo CRISP-DM. Vamos agora tratar das tarefas de mineração. Antes, porém, vamos fazer duas questões sobre o assunto:

HORA DE PRATICAR!

(Ano: 2014 Banca: CESPE Órgão: ANTAQ Prova: Analista Administrativo - Infraestrutura de TI) A respeito de Data Warehouse e Data Mining, julgue os itens subsecutivos.

Em um processo de descoberta do conhecimento, um Data Mining executado para atingir uma meta pode falhar nas classes de predição, de identificação, de classificação e de otimização.

Comentário: Vejam que a modelagem existe justamente para que o processo possa ser testado antes de ir para execução. Ainda assim, podemos ter falhas devido a uma mudança nos perfis dos dados. Por isso que o fluxo é interativo e incremental. Logo, a alternativa está correta!

Gabarito: C.

(Ano: 2017 Banca: FCC Órgão: TRT-11 Cargo: Técnico Judiciário de TI – Q. 47) Sistemas do tipo **I** e do tipo **II** realizam tarefas diferentes, porém complementares. O tipo **I** é adequado para atividades como indexação de dados, alocação de custos, análises de séries temporais e análises "*what-if*". Porém, a maioria dos sistemas do tipo **I** não tem a capacidade de realizar inferências indutivas, processo que permite chegar a conclusões genéricas a partir de exemplos específicos, que são uma característica nativa de sistemas do tipo **II**. Sistemas do tipo **I** fornecem uma visão multidimensional de dados, incluindo suporte a hierarquias. Essa visão de dados é uma forma natural de analisar negócios e organizações. Sistemas do tipo **II**, por outro lado, podem ajudar a detectar tendências, encontrar padrões e relações entre as informações disponíveis em bancos de dados. Os sistemas do tipo **II** podem encontrar informações ocultas nos dados disponíveis, mas é o gestor quem deve atribuir o valor de cada uma dessas descobertas para a organização.
Os sistemas do tipo **I** e **II** são, correta e respectivamente,
(A) OLAP e Data Warehouse.
(B) Data Warehouse e Data Mining.
(C) Banco de Dados Multidimensional e Banco de Dados Relacional.
(D) Data Mining e Data Warehouse.
(E) OLAP e Data Mining.

Comentário: Para responder à questão, precisamos estar seguros de alguns conceitos. O primeiro deles se refere à capacidade de análise de uma ferramenta OLAP. Vejam o texto que eu encontrei na internet que trata da relação de OLAP e Data mining:

"OLAP é a sigla para On-Line Analytical Processing. Refere-se a análises rápidas de dados multidimensionais compartilhados. OLAP e mineração de dados são coisas diferentes, porém complementares.

OLAP é adequado para atividades como indexação de dados, alocação de custos, análises de séries temporais e análises "what-if". Porém, a maioria dos sistemas de OLAP não têm a capacidade de realizar inferências indutivas além das análises preditivas limitadas a esses fatores.

A inferência indutiva, processo que permite chegar a conclusões genéricas a partir de exemplos específicos, é uma característica nativa de data mining. Talvez você já tenha visto referências a esse conceito, como “aprendizado de máquina”.

Sistemas de OLAP fornecem uma visão multidimensional de dados, incluindo suporte total a hierarquias. Essa visão de dados é uma forma natural de analisar negócios e organizações. Minerar dados, por outro lado, normalmente não conta com os conceitos de dimensões e hierarquias.

Data mining pode ajudar a detectar tendências como “propensão de uma pessoa a comprar” e “propensão de um cliente a interromper a assinatura do serviço” (o infame “churn“). Os sistemas OLAP podem, então, agregar e indexar estas probabilidades.”

Curiosidade: O que seria uma análise “what-if”? Na análise do tipo **What If**, o usuário final **introduz mudanças nas variáveis** ou **nas relações entre varáveis** e observa as mudanças resultantes nos valores de outras variáveis. Ou seja, trata de como uma mudança de uma variável afeta outras Ex: O que ocorre se reduzirmos a propaganda em 10%?

Gabarito: E.

TÉCNICAS DE PRÉ-PROCESSAMENTO

Nesta parte da aula vamos apresentar algumas técnicas que nos permitem fazer ajustes nos dados para facilitar a mineração, bem como para garantir que a síntese feita sobre os dados produza alguma informação consistente.

Os dados no mundo real são “sujos”, ou seja, podem, por exemplo, estarem **incompletos** com valores faltantes, atributos faltantes. Outra possibilidade é a existência de **ruídos**, que provocam erros ou outliers sobre os dados. Por fim, podemos verificar a consistência entre atributos, por exemplo, os atributos data de nascimento e idade precisam ser compatíveis. Uma pessoa que nasceu em 1983 não pode ter, em 2017, 50 anos. Vejamos as definições para os problemas que podem ocorrer com os dados:

Incompletude: a incompletude de uma base de dados pode ocorrer de várias formas; por exemplo, podem **faltar valores de um dado atributo**; pode **faltar um atributo** de interesse; ou pode **faltar um objeto** de interesse. Note, entretanto, que nem sempre a ausência de um atributo ou um objeto é percebida, a não ser quando um especialista no domínio do problema analisa a base e percebe a falta – por exemplo, um professor que identifica a ausência do nome de um aluno (objeto) ou um dia da semana (atributo) na lista de chamada.

Inconsistência: em bancos de dados, um dado inconsistente ocorre quando diferentes e conflitantes versões do mesmo dado aparecem em locais variados. Na área de mineração de dados, um dado inconsistente normalmente é aquele cujo valor **está fora do domínio do atributo ou apresenta uma grande discrepância em relação aos outros dados**. Por exemplo, a idade de um aluno que defendeu o doutorado deveria ser, ao menos, 25 anos.

Ruído: a palavra ruído possui diversos significados, dependendo do contexto. Por exemplo, em vídeo, um ruído é aquele chuvisco na imagem e, em rádio, é aquela interferência no sinal de áudio. Entretanto, a noção de ruído em mineração de dados está

mais próxima do conceito de ruído em **estatística** (variações inexplicáveis em uma amostra) e processamento de sinais (variações indesejadas e normalmente inexplicáveis em um sinal). Um dado ruidoso é aquele que apresenta alguma variação em relação ao seu valor sem ruído e, portanto, ruídos na base de dados podem levar a inconsistências. Cabe ressaltar que, dependendo do nível de ruído, nem sempre é possível saber se ele está ou não presente em um dado.

De posse dessas definições, podemos organizar os tipos de problemas da seguinte forma;

Figura 1 - Principais problemas com os dados

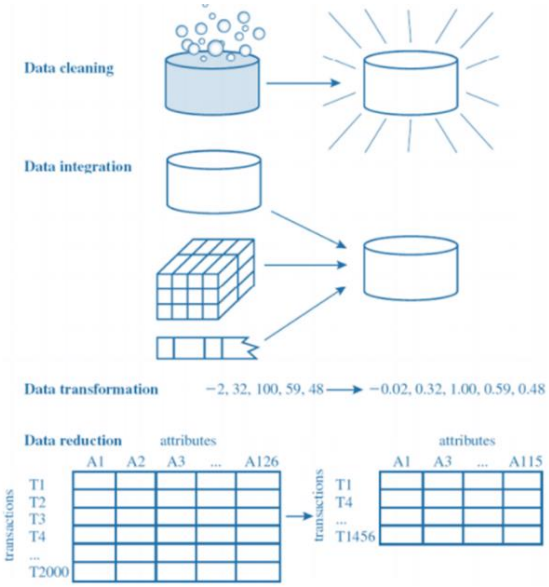
Técnicas de pré-processamento e transformação de dados são aplicadas para **aumentar a qualidade** e o **poder de expressão dos dados** a serem minerados. Dados sem qualidade podem gerar uma mineração sem qualidade, que por consequência vão levar a decisões sem qualidade. Se os dados estiverem duplicados ou faltantes podemos gerar cálculos estatísticos incorretos.

A melhor maneira de pré-processar os dados depende de três fatores centrais: os problemas (incompletude, inconsistência e ruído) existentes na base bruta; quais respostas pretendem-se obter das bases, ou seja, qual problema deve ser resolvido; e como operam as técnicas de mineração de dados que serão empregadas. Esses três fatores quase sempre estão inter-relacionados.

Esta fase tende a consumir uma parte significativa do tempo dedicado ao processo de KDD. E para avaliar se a qualidade dos dados está de acordo com a necessidade das análises de mineração, precisamos avaliar alguns aspectos dos dados, como: acurácia, completude, consistência, se eles estão corretos em relação ao tempo, à confiabilidade, ao grau de agregação de valor, à sua capacidade de interpretação e à acessibilidade.

Neste sentido, algumas tarefas de pré-processamento podem ser executadas. Antes de apresentar as tarefes, gostaria de fazer duas considerações. Primeiramente perceba que essas tarefas têm uma correlação com as atividades de transformação do processo de ETL. Segundo, não confunda essas tarefas (de pré-processamento) com as tarefas de mineração que veremos a seguir.

Limpeza dos Dados – Preenche valores faltantes, suaviza dados ruidosos, identifica ou remove "outliers" e resolve inconsistências.
Integração – Dados de origens diferentes devem ser integrados.
Transformação – Normalização e agregação dos dados.
Redução - Tenta reduzir o volume de dados sem provocar grandes alterações no resultado. Vamos falar um pouco mais sobre isso abaixo.
Discretização – Faz parte do processo de redução, mas tem papel importante, especialmente com dados numéricos. Visa estabelecer valores discretos para variáveis contínuas.



(Ministério da Economia – Especialista em Ciência de Dados - 2020) Acerca de conceitos, premissas e aplicações de big data, julgue os itens subsequentes.

O objetivo das técnicas de pré-processamento de dados é preparar os dados brutos para serem analisados sem erros de incompletudes, inconsistências e ruídos

Comentários: Segundo HAN e KAMBER, o pré-processamento pode melhorar a qualidade dos dados, melhorando assim a acurácia e eficiência dos processos de mineração subsequentes. Sua aplicação é necessária, pois as bases de dados em geral são muito grandes e contém registros que comprometem a qualidade dos dados, como por exemplo, registros inconsistentes, falta de informação (registros faltantes), registros duplicados, outliers (valores discrepantes), assimetria, transformação entre outros.

Entre as técnicas de pré-processamento temos:
- Limpeza dos dados (data cleaning)
- Integração de dados (data integration)
- Transformação de dados (data transformation)
- Redução de dados (data reduction)

Gabarito: CERTO.

Redução de dados

É intuitivo pensar que, quanto maior a quantidade de objetos e atributos, mais informações estão disponíveis para o algoritmo de mineração de dados. Entretanto, o aumento do número de objetos e da dimensão do espaço (número de atributos na base) pode fazer com que os dados disponíveis se tornem esparsos e as medidas matemáticas usadas na análise tornem-se numericamente instáveis. Além disso, uma quantidade muito grande de objetos e atributos pode tornar o processamento dos algoritmos de mineração muito complexo, assim como os modelos gerados.

Nesses casos, as técnicas de redução de dados podem ser aplicadas tanto para reduzir a quantidade de objetos da base quanto para reduzir a quantidade de atributos que os descrevem (dimensionalidade). Dentre os métodos de redução de dados destacam-se:

Seleção de atributos (ou características): efetua uma redução de dimensionalidade na qual atributos irrelevantes, pouco relevantes ou redundantes são detectados e removidos.

Compressão de atributos: também efetua uma redução da dimensionalidade, mas empregando algoritmos de codificação ou transformação de dados (atributos), em vez de seleção.

Redução no número de dados: neste método, os dados são removidos, substituídos ou estimados por representações menores (mais simples), como modelos paramétricos (que armazenam apenas os parâmetros do modelo em vez dos dados) e os métodos não paramétricos, como agrupamento, amostragem e histogramas.

Discretização: os valores de atributos são substituídos por intervalos ou níveis conceituais mais elevados, reduzindo a quantidade final de atributos.

Agora que já conseguimos diferenciar algumas tarefas de pré-processamento, observe a tabela para entender um pouco mais esses conceitos.

Tabela 1 - Palavras-chave das etapas de pré-processamento

Característica	Palavras-chave
Limpeza dos dados	Imputar valores ausentes, suavizar ruídos, identificar valores discrepantes (outliers) e corrigir inconsistências.
Integração dos dados	Resolver conflitos e redundância
Redução dos dados	Compressão de atributos e redução do número de dados.
Transformação dos dados	Padronização e Normalização
Discretização	Reduz a quantidade de valores de um dado atributo contínuo, facilitando, em muitos casos, o processo de mineração.

TAREFAS DE MINERAÇÃO

As funcionalidades de mineração de dados são usadas para especificar os tipos de padrões ou conhecimentos encontrados **nas tarefas de mineração de dados**. As funcionalidades incluem caracterização e discriminação; a mineração de padrões frequentes, associações e correlações; classificação e regressão; análise de cluster; e detecção de outliers. À medida que novos tipos de dados, novos aplicativos e novas demandas de análise continuarem surgindo, não há dúvida de que veremos mais e mais novas tarefas de mineração de dados.

Muitas das tarefas tentam encontrar padrões úteis sobre os dados. Esses padrões podem ser encontrados usando o índice de correlação. A **correlação** mede **o grau de relacionamento entre duas variáveis**. Quando **positiva** a correlação indica que as variáveis caminham juntas, ou seja, se uma variável cresce a outra também deve seguir a mesma direção. Por outro lado, se a correlação for **negativa**, o aumento de uma variável implica na diminuição da outra. Essa medida, em geral é avaliada de forma **síncrona**. Ou seja, você mede as duas variáveis no mesmo instante e verifica a existência de uma correlação.

Entretanto, é possível que variações em uma variável influenciem uma outra variável no instante posterior. Perceba que, neste caso, as variáveis serão medidas em momentos distintos. Esse tipo de correlação é chamado de **assíncrona**. Um exemplo simples seria a influência do índice da bolsa de Nova Iorque na bolsa de Tóquio. Por não estarem abertas simultaneamente, a medição dos valores e do grau de correlação entre eles ocorrem em momentos diferentes, ou seja, são assíncronos.

(Ministério da Economia – Especialista em Ciência de Dados - 2020) No que se refere à mineração de dados, julgue os itens a seguir.

A correlação assíncrona pode indicar um alto coeficiente de similaridade entre dois eventos, mesmo que eles tenham iniciado em momentos distintos.

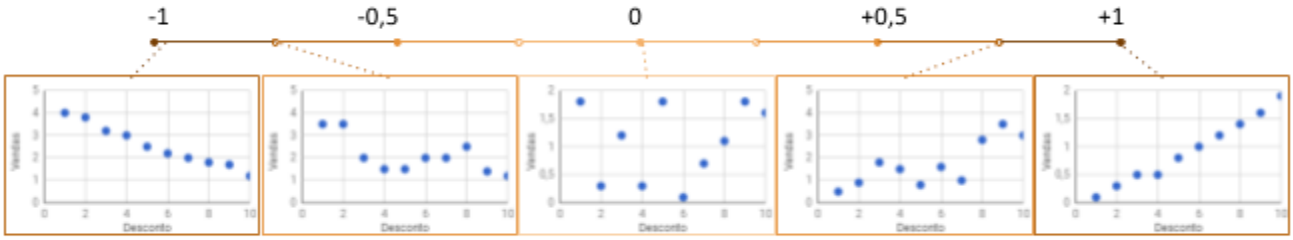
Comentários: A Correlação mede a força ou grau de relacionamento entre duas variáveis. Geralmente as variações acontecem simultaneamente, ou seja, são consideradas correlações síncronas. Contudo, é possível que a alteração de uma variável afete outra em um momento futuro, essas são chamadas correlações assíncronas.

A Correlação pode ser classificada, quanto ao sentido, em positiva ou negativa. Uma correlação positiva indica que à medida que a variável x aumenta implica que a variável y também aumenta, se a variável x diminui isso também ocorrerá com a variável y. A Correlação será negativa se o aumento de uma variável implica na diminuição da outra.

Quando a correlação assíncrona é medida, ela pode indicar um alto grau de similaridade entre eventos que iniciaram em momentos distintos.

Gabarito: CERTO.

O indicativo que diz se há ou não há correlação, ou se ela é **positiva** ou **negativa**, é o coeficiente de correlação, que é um número que varia de -1 a +1, como na régua a seguir. Observe que quanto mais próximo de 1 for o valor do coeficiente, sendo positivo ou negativo, mais forte é a evidência de que há uma relação entre as duas variáveis.



(Ministério da Economia – Especialista em Ciência de Dados - 2020) Julgue os seguintes itens, a respeito de big data

Em se tratando da técnica de correlação, utiliza-se uma escala de 1 a -1 para indicar o grau de similaridade entre duas variáveis distintas.

_____

Comentários: O indicativo que diz se há ou não há correlação, ou se ela é positiva ou negativa, é o coeficiente de correlação, que é um número que varia de -1 a +1.

Gabarito: CERTO.

Vamos falar agora das tarefas de mineração, começando pela regra de associação.

REGRAS DE ASSOCIAÇÃO

As regras de associação relacionam a presença de um conjunto de itens com outra faixa de valores de outro conjunto de variáveis. Podemos pensar nos seguintes exemplos: 1. Quando uma mulher compra uma bolsa em uma loja, ela está propensa a comprar sapatos (na mesma loja) e 2. Uma imagem de raio X contendo as características a e b provavelmente exibirá também a característica c (o mesmo raio-x). Veja as figuras abaixo que ilustram esses exemplos:

Uma regra de associação é um padrão da forma X → Y, onde X e Y são conjuntos de valores. O seguinte padrão “clientes que compram pão também compram leite” representa uma regra de associação que reflete um padrão de comportamento dos clientes do

supermercado. Descobrir regras de associação entre produtos comprados por clientes numa mesma compra pode ser útil para melhorar a organização das prateleiras, facilitar (ou dificultar) as compras do usuário ou induzi-lo a comprar mais.

Os autores definem os conceitos de lado da mão direita e lado da mão esquerda para ilustrar essa ideia de compra casada. É como se eu estivesse propenso a consumir os dois produtos. A união entre o lado da mão esquerda e o lado da mão direita gera outra definição conhecida como conjunto-item (o conjunto de todos os itens comprados pelos clientes). Observe a figura abaixo com o conjunto-item formado por picanha (Friboi é claro!) e carvão!

Suporte: % (LME U LMD)

Confiança: Suporte (LME U LMD)/ Suporte (LME)

Para que uma regra de associação seja do interesse de um pesquisador de dados, a regra precisa satisfazer algumas medidas. O suporte que define quão frequente a regra acontece no banco de dados e a confiança que é a força da regra. Vamos detalhar um pouco mais essas definições.

O **Suporte** é uma **medida objetiva** para avaliar o **interesse** de uma **regra de associação.** Representa **a porcentagem de transações (%)** de um banco de dados de transações **onde a regra se verifica.** A medida de suporte responde a seguinte questão: quão frequente a regra acontece no banco de dados?

A **Confiança** é outra medida **objetiva** para **regras de associação** que mede o **grau de certeza** de uma associação. Em termos estatísticos, trata-se simplesmente da **probabilidade condicional P (Y | X)**, isto é, a porcentagem de transações contendo os itens de X que também contêm os itens de Y. Confiança é a força da regra

Vamos fazer uma questão sobre o assunto. Desta vez, a banca em questão é a FGV.

(Ano: 2008 Banca: FGV Órgão: Senado Federal Cargo: Analista de Sistemas) Considerando as diferentes técnicas de mineração de dados, NÃO é correto afirmar que:

A) em Regras de Associação, confiança refere-se a quantas vezes uma regra de associação se verifica no conjunto de dados analisado.
B) correlação canônica e análise múltipla de discriminante são técnicas utilizadas para análise multivariada.

C) na análise de grupamentos, medidas de correlação, medidas de distância e medidas de associação são alguns dos métodos utilizados para medir a semelhança entre objetos.
D) a classificação é considerada um exemplo de aprendizado supervisionado, enquanto o agrupamento é considerado exemplo de aprendizado não supervisionado.
E) regressão é uma aplicação especial da regra de classificação, onde a regra é considerada uma função sobre variáveis, mapeando-as em uma classe destino.

_____

Comentários: Vejam que a alternativa A apresenta um conceito incorreto, quem se refere à quantidade de vezes que uma regra se verifica é o suporte e não a confiança. Desta forma, como a questão pede a alternativa incorreta, essa é a nossa resposta. A demais alternativas estão corretas, falaremos sobre elas nas próximas páginas.

Gabarito: A.

(Ministério da Economia – Especialista em Ciência de Dados - 2020) No que se refere à mineração de dados, julgue os itens a seguir.

A técnica de associação é utilizada para indicar um grau de afinidade entre registros de eventos diferentes, para permitir o processo de data mining

_____

Comentários: Regra de associação está associada a palavra-chave coocorrência. Assim, estamos procurando por eventos ou informações que tenham alto grau de afinidade.

Gabarito: CERTO.

(Ministério da Economia – Especialista em Ciência de Dados - 2020) Julgue os seguintes itens, a respeito de big data

Tratando-se de aprendizagem de máquina, o fator de confiança para as evidências varia de -1 a 1 para representar a certeza do fato.

_____

Comentários: O fator de confiança está associado ao percentual de um conjunto de transações em que uma regra de associação se verifica dado que o conjunto de itens a priori (LME) está presente.

Gabarito: ERRADO.

O problema de regras de associação pode ser decomposto em três passos principais:

1. Gerar todas as combinações de itens;

2. Descobrir conjuntos de itens: Este passo consiste em gerar um conjunto com todas as combinações de itens obedecendo a um limiar, chamado **suporte mínimo**. As combinações que satisfazem esta condição são chamadas de

conjunto de itens grandes, enquanto os que não satisfazem são chamados de **conjunto de itens pequenos**;

3. Gerar as regras de associação para a base de dados: Após o conjunto de itens finais ter sido produzido, deve-se gerar as regras de associação de um conjunto de itens Y = I1,I2,,...,Ik, sendo k ≥ 2. O antecedente da regra será um conjunto X de Y tal que, X possua k-1 itens, e o consequente seja Y – X. Para verificar a validade de uma regra, a confiança da regra (suporte(Y) / suporte(X)) deve satisfazer o valor mínimo de confiança informado.

Uma pergunta pode ser relevante neste momento: **como descobrir todos os conjuntos de itens grandes?**

Temos que verificar duas propriedades:

1. **Fechamento por baixo,** ou seja, um itemset grande também deve ser grande (desta forma cada subconjunto de um itemset excede o suporte mínimo exigido).

2. **Antimonotonicidade** um superconjunto de um itemset pequeno também é pequeno (implicando que ele não tem suporte suficiente). Sendo assim, quando se descobre um itemset pequeno, então qualquer extensão deste itemset será pequeno.

CLASSIFICAÇÃO

Parece ser um imperativo humano. A fim de compreender e comunicar sobre o mundo que estamos constantemente a classificar, categorizar e classificar. Dividimos as coisas vivas em filos, espécies e gênero; matéria em elementos; cães em raças, as pessoas em raças. Os objetos a serem classificados são geralmente representados por registros em um banco de dados ou um arquivo, e o ato de classificação consiste em adicionar uma nova coluna com um código de classe de algum tipo.

Uma das tarefas mais comuns dentro de mineração de dados consiste em examinar as características de um objeto recém-apresentado e atribuí-lo a um dos conjuntos predefinidos de classes. A tarefa de classificação é caracterizada por uma definição das classes (1), e conjunto dados para aprendizado (2) pré-classificados.

Uma definição mais formal para a classificação é a tarefa de aprendizado de uma função alvo f que mapeia cada atributo de um conjunto x para um rótulo de classe predefinido y. Essa descrição foi dada por Tan em seu livro de mineração e pode ser observada na figura abaixo:

O modelo construído baseia-se na análise prévia de um conjunto de dados de amostragem ou dados de treinamento, contendo objetos corretamente classificados. Por exemplo, suponha que o gerente do supermercado está interessado em descobrir que tipo de características classificam seus clientes em “bom comprador” ou “mau comprador”. Um modelo de classificação poderia incluir a seguinte regra: “Clientes da faixa econômica B, com idade entre 50 e 60 são maus compradores”.

Vejamos uma questão do CESPE sobre esse assunto:

HORA DE PRATICAR!

(Ano: 2015 Banca: CESPE Órgão: MEC Prova: Administrador de Banco de Dados) Julgue o item seguinte, referente a data mining.

O conhecimento obtido no processo de data mining pode ser classificado como uma regra de associação quando, em um conjunto de eventos, há uma hierarquia de tuplas sequenciais.

_____

Comentário: Vejam que a tarefa de classificação tem um objetivo diferente da regra de associação. Os eventos que ocorrem em conjunto, como as compras em um supermercado, são descritos por uma regra de associação.

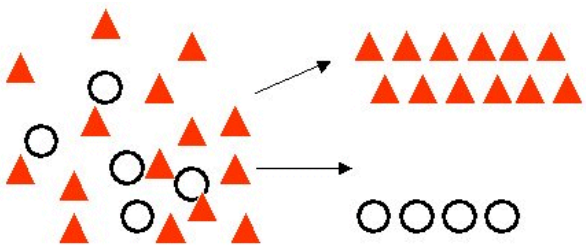
A classificação reconhece modelos que descrevem o grupo ao qual o item pertence por meio do exame dos itens já classificados e pela inferência de um conjunto de regras.

Por exemplo: empresas de operadoras de cartões de crédito e companhias telefônicas preocupam-se com a perda de clientes regulares, a classificação pode ajudar a descobrir as características de clientes que provavelmente virão abandoná-las e oferecer um modelo para ajudar os gerentes a prever quem são, de modo que se elabore antecipadamente campanhas especiais para reter esses clientes. Assim, observamos que a alternativa está incorreta.

Gabarito: E.

São técnicas usualmente empregadas em tarefas de classificação, árvores de decisão e redes neurais. Boa parte dos métodos de classificação utilizam técnicas estatísticas e de aprendizado de máquina. Segundo o Navathe, classificação é o processo de encontrar um conjunto de modelos (funções) que descrevem e distinguem classes ou conceitos.

Tem o propósito de utilizar o modelo para predizer a classe de objetos que ainda não foram classificados. Utiliza um aprendizado supervisionado para separar classes em grupos distintos. Vejam um exemplo na figura abaixo:



Na classificação, o objetivo é a construção de um modelo que possa ser aplicado a dados não classificados e classificá-los. São exemplos de tarefas de classificação que

foram abordados por meio de técnicas de mineração de dados: classificação de pedido de crédito como baixo, médio ou alto risco, escolher conteúdo a ser exibido em uma página Web, determinar quais os números de telefone correspondem a máquinas de fax, descobrir sinistros fraudulentos e atribuir códigos da indústria e denominações de emprego com base nas descrições de texto livre.

Em todos os exemplos, há **um número limitado de classes**, e espera-se ser capaz de atribuir qualquer registo em um ou outra. As **árvores de decisão** e técnicas semelhantes são bem adaptadas para a classificação. **Rede neural** e análise de links também são úteis para a classificação de certas circunstâncias. Vejam na figura a seguir um fluxo que mostra o funcionamento de um algoritmo de classificação:

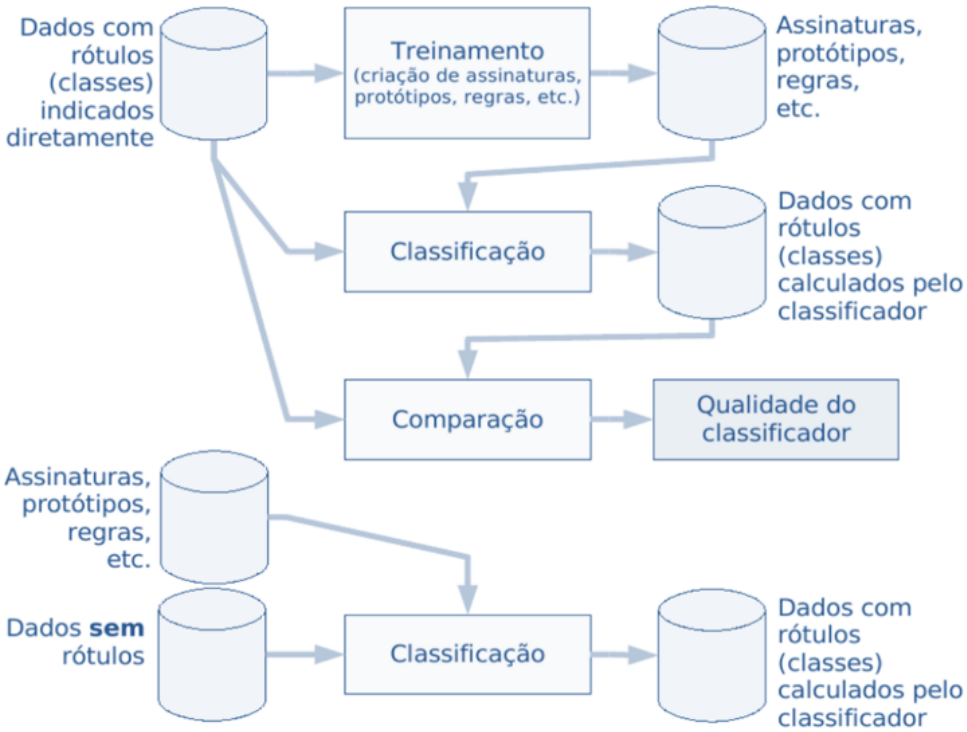


Figura 2 - Processo de construção de um classificador

Antes de seguirmos para a próxima tarefa, vejamos uma questão do CESPE sobre esse assunto:

HORA DE PRATICAR!

(Ano: 2016 Banca: CESPE Órgão: FUNPRESP-JUD Prova: Analista - Tecnologia da Informação) Julgue o item subsecutivo, referente às tecnologias de bancos de dados.

Em Data Mining, as árvores de decisão podem ser usadas com sistemas de classificação para atribuir informação de tipo.

_____

Comentário: A questão mostra uma das técnicas que podem ser usadas para implementação da tarefa de classificação. Vimos na parte teórica da aula que essa afirmação está correta.

Gabarito: C.

AGRUPAMENTO (CLUSTERING)

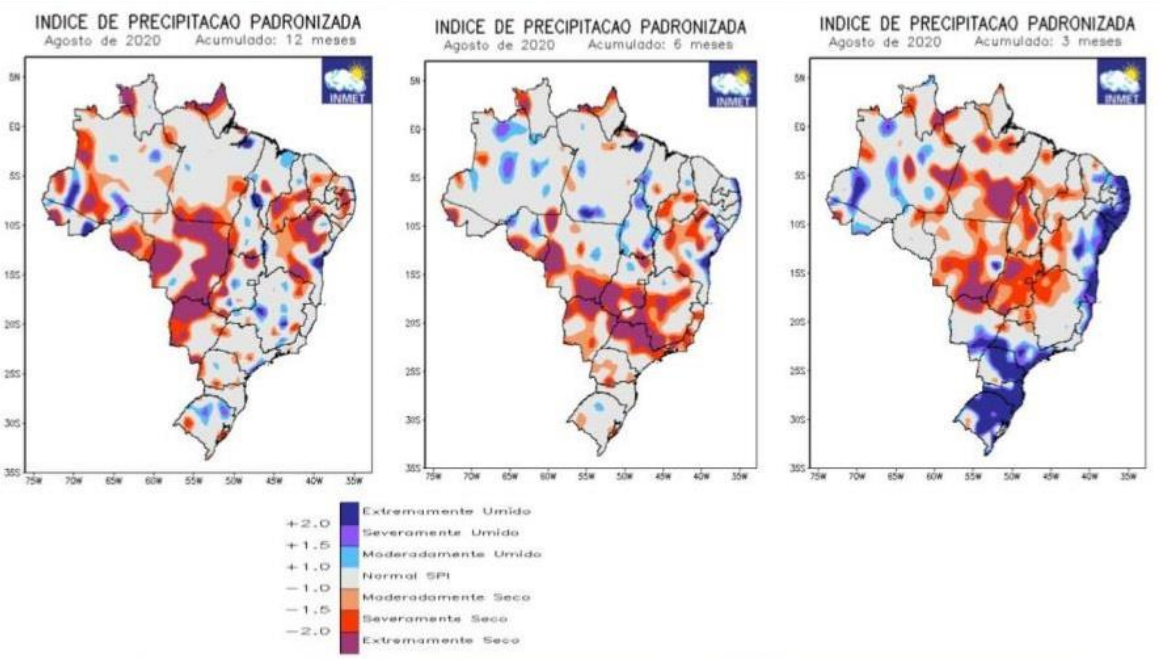
Diferentemente da classificação e predição, onde os dados de treinamento estão devidamente classificados e as etiquetas das classes são conhecidas, a análise de clusters trabalha sobre dados onde as etiquetas das classes não estão definidas. Então, qual o objetivo do agrupamento?

Encontrar **grupos de objetos** tal que objetos em um grupo são **similares** (ou relacionados) uns aos outros e diferentes de (ou não relacionados) a objetos em outros grupos.

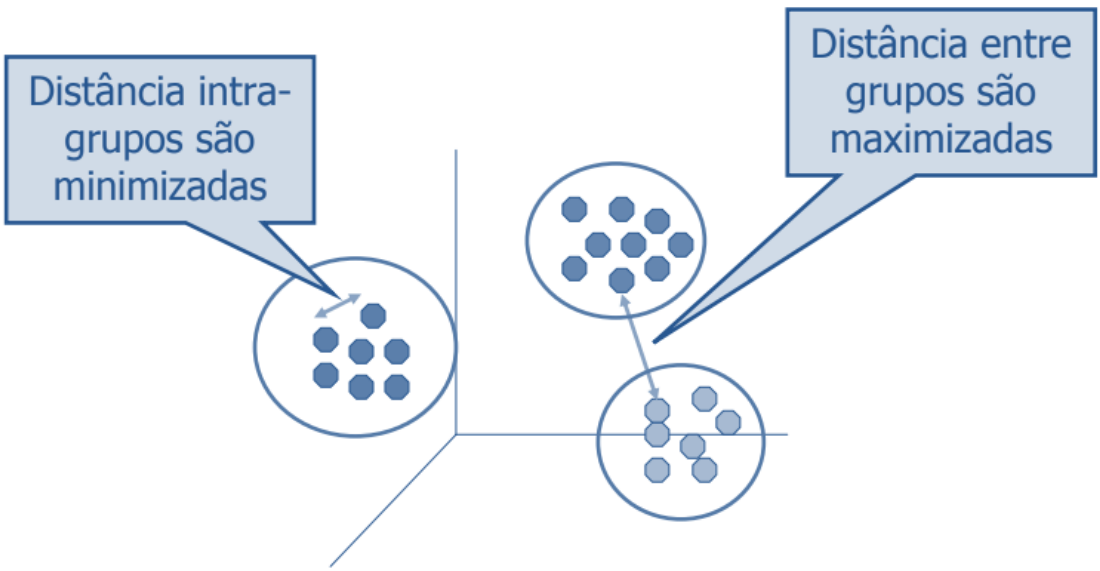
A técnica consiste em identificar **agrupamentos de objetos.** O agrupamento trabalha sobre dados onde as etiquetas das classes **não estão definidas previamente**. Por isso, podemos dizer que a clusterização é uma tarefa de **aprendizado não supervisionado**. **Perceba que a saída dos algoritmos que executam o trabalho de agrupar os dados são conjuntos de dados com objetos similares.**

Ela pode ser usada para facilitar a **compreensão dos dados**, agrupando documentos relacionados para pesquisa, genes e proteínas que tenham funcionalidades similares, ou estoques com flutuações de preço similar. É possível ainda usar o agrupamento para sumarizar em um conjunto de propriedades cada um do conjunto de dados.

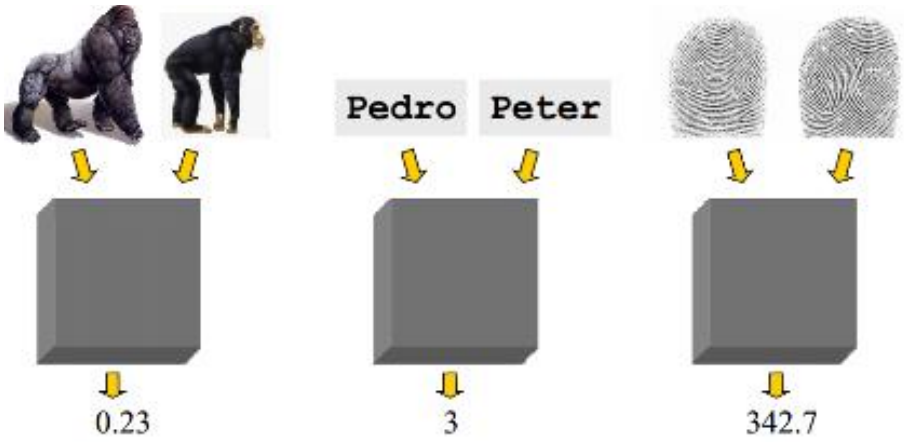
Pense que você tem um conjunto de dados sobre precipitação de água de chuva, sabemos que no Nordeste temos um ambiente mais seco e árido, enquanto na Amazônia temos um ambiente mais úmido com alta precipitação de chuvas. Podemos agrupar o mapa brasileiro de acordo com a precipitação de chuvas em cada local. Veja que os índices pluviométricos de cada região são similares.



Ok! Queremos achar grupos naturais nos dados onde as informações em um mesmo grupo são semelhantes entre si e dados de grupos diferentes são diferentes entre si. Mas, como medir essa semelhança? Existem algumas métricas para calcular **as distâncias e as dissimilaridades entre os itens de dados**.



As medidas de distância podem ser calculadas entre dois objetos. Sejam O1 e O2 dois objetos de um universo de possíveis objetos. A distância (dissimilaridade) entre O1 e O2 é um número real denotado por D (O1, O2). Observem a figura abaixo para entender melhor o conceito:



Algumas propriedades podem ser analisadas nas medidas de distância. A **simetria** (D (A, B) = D (B,A)), caso contrário você poderia afirmar que “Alex parece com Bob, mas Bob não parece com Alex”. A **constância de autos simetria** (D (A, A) = 0), caso contrário você poderia afirmar que “Alex parece mais com Bob, do que o próprio Bob”.

A **positividade** (D(A,B) = 0 ←→ A = B) caso contrário existiriam objetos no seu mundo que são diferentes, mas você não consegue diferenciá-los. E a **desigualdade triangular** (D (A,B) ≤ D(A,C) + D (B,C)) caso contrário você poderia afirmar que “Alex é parecido com Bob, e Alex é parecido com Carl, mas Bob não se parece com Carl”.

O CESPE já cobrou isso em provas anteriores, vamos ver como:

HORA DE PRATICAR!

(Ano: 2017 Banca: CESPE Órgão: SEDF Prova: Analista de Gestão Educacional - Tecnologia da Informação) Com relação a data mining e data warehouse, julgue o item que se segue.

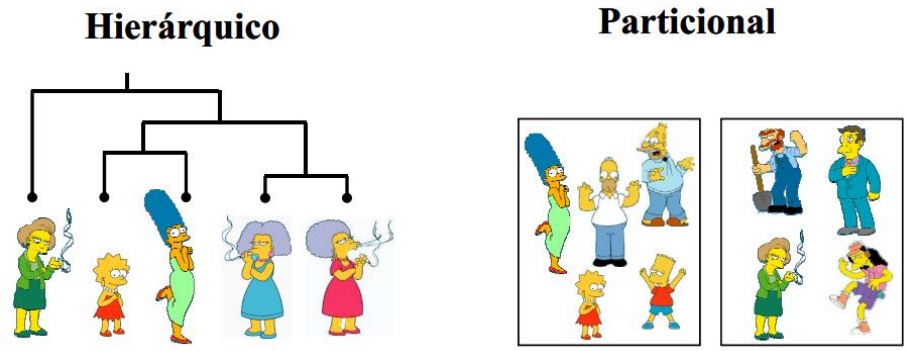
Agrupar registros em grupos, de modo que os registros em um grupo sejam semelhantes entre si e diferentes dos registros em outros grupos é uma maneira de descrever conhecimento descoberto durante processos de mineração de dados.

Comentário: Vejam que a questão apresenta uma definição coerente, a aglomeração (clustering) funciona de maneira semelhante à classificação quando ainda não foram definidos grupos. Uma ferramenta de data mining descobrirá diferentes agrupamentos dentro da massa de dados. Por exemplo, ao encontrar grupos de afinidades para cartões bancários ou ao dividir o banco de dados em categorias de clientes com base na demografia e em investimentos pessoais.
Neste caso temos que lembrar que o interesse da **aglomeração é segmentar uma amostra** em classe pré-definida. Não estamos tentando classificar novos valores. A alternativa, portanto, está correta!

Gabarito: CERTO.

Podemos usar alguns tipos de algoritmos para agrupamentos. Algoritmos **particionais** que objetivam construir diversas partições e avaliá-las com algum critério, ele divide objetos de dados em subconjuntos sem sobreposição (grupos) tal que cada objeto de dados está em

exatamente um subconjunto. Algoritmos **hierárquicos** que criam uma decomposição hierárquica de um conjunto de objetos utilizando algum critério. Um conjunto de grupos aninhados é organizado como uma árvore hierárquica. Vejam o exemplo abaixo:



Vejamos uma questão sobre o assunto ...

HORA DE PRATICAR!

(Ministério da Economia – Especialista em Ciência de Dados - 2020) No que se refere à mineração de dados, julgue os itens a seguir.

Na análise hierárquica de agrupamentos, é possível realocar um elemento que tenha sido alocado incorretamente no início do processo

_____

Comentários: Os métodos hierárquicos da análise de cluster tem como principal característica a criação de um algoritmo capaz de fornecer mais de um tipo de partição dos dados. Ele gera vários agrupamentos possíveis, onde um cluster pode ser mesclado a outro em determinado passo do algoritmo. Esses métodos não exigem que já se tenha um número inicial de clusters e são considerados inflexíveis uma vez que **não se pode trocar um elemento de grupo**.

Gabarito: ERRADO.

Por exemplo, poderíamos aplicar análise de clusters sobre o banco de dados de um supermercado a fim de identificar grupos homogêneos de clientes. Clientes residentes em determinados pontos da cidade costumam vir ao supermercado aos domingos. Enquanto clientes residentes em outros pontos da cidade costumam fazer suas compras às segundas-feiras.

HORA DE PRATICAR!

(CESPE - 2013 - MPU - Analista - Suporte e Infraestrutura) Julgue os próximos itens, acerca de sistemas de suporte à decisão.

Em se tratando de mineração de dados, a técnica de agrupamento (clustering) permite a descoberta de dados por faixa de valores, por meio do exame de alguns atributos das entidades envolvidas.

Comentário: Segundo Navathe, "O objetivo do agrupamento é colocar registros em grupos, de modo que os registros em um grupo sejam semelhantes uns aos outros e diferentes dos registros em outros grupos. Os grupos costumam ser disjuntos."

Gabarito: CERTO.

ABORDAGEM PARA OUTROS PROBLEMAS DE MINERAÇÃO

Análise de padrões sequenciais - Um padrão sequencial é uma expressão da forma < i1;.....; in >, onde cada i é um conjunto de itens. A ordem em que estão alinhados estes conjuntos reflete a ordem cronológica em que aconteceram os fatos representados por estes conjuntos. Assim, por exemplo, a sequência < {carro}, {pneu, toca-fitas} > representa o padrão “Clientes que compram carro, **tempos depois compram** pneu e toca-fitas de carro”. Descobrir tais padrões sequenciais em dados temporais pode ser **útil** em **campanhas de marketing**, por exemplo.

Outro exemplo de padrão sequencial interessante: você compra o material do Estratégia Concursos®, estuda com afinco e cumpre todas as metas de estudos e se torna servidor público. Esse é um padrão que estamos verificando com frequência aqui no Estratégia. Vejamos uma questão sobre esse assunto.

(Ministério da Economia – Especialista em Ciência de Dados - 2020) Julgue os seguintes itens, a respeito de big data

O objetivo da técnica de sequência de tempo é identificar a ocorrência de dois eventos diferentes no mesmo momento.

Comentários: A técnica de padrões sequenciais é usada para descobrir eventos que ocorrem um após o outro no tempo. O objetivo é identificar ações ou eventos encadeados.

Gabarito: ERRADO.

Análise de Padrões em Séries Temporais - O preço de fechamento de uma ação ou de um fundo de investimentos é um evento que ocorre a cada dia da semana para cada fundo

ou ação. Sequências desses valores são exemplos de uma serie temporal. Séries temporais são sequências de eventos, cada evento pode ser um tipo fixo dado uma transação.

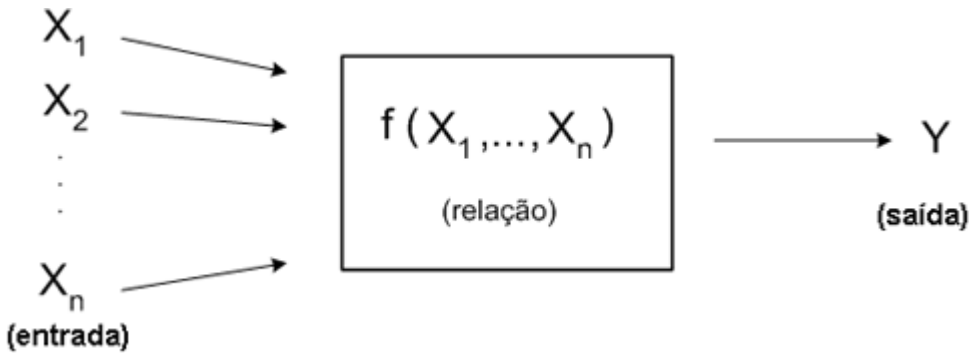
Uma série temporal é uma coleção de observações feitas sequencialmente ao longo do tempo. Em séries temporais, a ordem dos dados é fundamental. Uma característica muito importante deste tipo de dados é que as observações vizinhas são dependentes e o interesse é analisar e modelar esta dependência.

Predição – Consultando o dicionário, encontramos a seguinte definição: dizer antecipadamente o que vai acontecer, seja por meio de regras certas, pretensa adivinhação ou conjetura. Em algumas aplicações, o usuário está mais interessado em predizer alguns valores ausentes em seus dados, em vez de descobrir classes de objetos. Isto ocorre, sobretudo, quando os valores que faltam são numéricos. Neste caso, a tarefa de mineração é denominada Predição.

Análise de Outliers - Um banco de dados pode conter dados que não apresentam o mesmo comportamento padrão da maioria. Estes dados são denominados outliers (exceções). Muitos métodos de mineração descartam estes outliers como sendo ruído indesejado. Entretanto, em algumas aplicações, tais como detecção de fraudes, estes eventos raros podem ser mais interessantes do que eventos que ocorrem regularmente. Por exemplo, podemos detectar o uso fraudulento de cartões de crédito ao descobrir que certos clientes efetuaram compras de valor extremamente alto, fora de seu padrão habitual de gastos.

Veja que você pode descobrir fraudes por análise de outleirs ou classificação. No primeiro, você conhece o padrão dos dados ou transações comuns ao banco de dados e algo que não se adapte a esse padrão é um ponto fora da curva. No segundo você treina um modelo para que ele reconheça padrões de fraude.

Regressão - Regressão é uma aplicação especial da regra de classificação. Se uma regra de classificação é considerada uma função sobre variáveis que as mapeia em uma classe destino, a regra é chamada regressão. Uma aplicação de regressão ocorre quando, em vez de mapear uma tupla de dados de uma relação para uma classe específica, o valor da variável é previsto baseado naquela tupla.



Quando: Y = f(x1, x2, ..., xn). Uma função f é linear no domínio das variáveis xi, o processo de derivar f de um dado conjunto de tuplas para <x1, x2, ... Xn, y> é chamado **regressão linear**. Modelos de Regressão são construídos com os objetivos:

INDO MAIS FUNDO!

Descrição – É utilizada uma equação para sumarizar ou descrever o relacionamento de um conjunto de dados, onde a análise de regressão pode ser empregada para ajustar a equação.

Predição - Uma vez que esperamos que grande parte da variação da variável de saída seja explicada pelas variáveis de entrada, podemos utilizar o modelo para obter valores de Y correspondentes a valores de X que não estavam entre os dados. Esse procedimento é chamado de predição e, em geral, usamos valores de X que estão dentro do intervalo de variação estudado. A utilização de valores fora desse intervalo recebe o nome de extrapolação e deve ser usada com muito cuidado, pois, o modelo adotado pode não ser correto fora do intervalo estudado. Acredita-se que a predição seja a aplicação comum dos modelos de regressão;

Controle – É muito utilizado na regressão com o objetivo de controlar a variável de interesse em faixa de valores pré-fixadas. Destaca-se, quando a equação de regressão é empregada, a relação existente entre a variável de interesse e as variáveis utilizadas para seu controle seja do tipo causa e efeito.

Seleção de variáveis - Frequentemente, não se tem ideia de quais são as variáveis que afetam significativamente a variação de Y. Para responder a esse tipo de questão, estudos são realizados com muitas variáveis. A análise de regressão pode auxiliar no processo de seleção de variáveis eliminando aquelas cuja contribuição não seja importante;

Estimação (de parâmetros) - Dado um modelo e um conjunto de dados referente às variáveis resposta e preditoras, estimar parâmetros ou ajustar um modelo aos dados significa obter valores ou estimativas para os parâmetros, por algum processo, tendo por base o modelo e os dados observados;

Inferência - O ajuste de um modelo de regressão em geral tem por objetivos básicos, além de estimar os parâmetros, realizar inferências sobre eles, tais como, testes de hipóteses e intervalos de confiança.

HORA DE PRATICAR!

(Ano: 2015 Banca: CESPE Órgão: MEC Prova: Administrador de Dados) Acerca de data warehouse (DW), Business Intelligence (BI) e data mining, julgue o item que se segue.

Situação hipotética: Após o período de inscrição para o vestibular de determinada universidade pública, foram reunidas informações acerca do perfil dos candidatos,

cursos inscritos e concorrências. Ademais, que, por meio das soluções de BI e DW que integram outros sistemas, foram realizadas análises para a detecção de relacionamentos sistemáticos entre as informações registradas. Assertiva: Nessa situação, tais análises podem ser consideradas como data mining, pois agregam valor às decisões do MEC e sugerem tendências, como, por exemplo, o aumento no número de escolas privadas e a escolha de determinado curso superior.

_____

Comentário: Observem que a afirmação está correta e de acordo com o que vimos até aqui. A mineração de dados ajuda a identificar tendências sobre os dados. Essas tarefas são conhecidas como preditivas.

Gabarito: C.

(Ministério da Economia – Especialista em Ciência de Dados - 2020) Julgue os seguintes itens, a respeito de big data

A análise de regressão em mineração de dados tem como objetivos a sumarização, a predição, o controle e a estimação.

_____

Comentários: Uma aplicação de regressão ocorre quando, em vez de mapear uma tupla de dados de uma relação para uma classe específica, o valor da variável é previsto baseado naquela tupla. A análise de regressão consiste na realização de uma análise estatística com o objetivo de verificar a existência de uma relação funcional entre uma variável dependente com uma ou mais variáveis independentes. De maneira geral, a análise de regressão tem como objetivos a sumarização, a predição, o controle, a estimação, a seleção de variáveis e a inferência.

Gabarito: CERTO.

CONCEITOS COMPLEMENTARES

Para concluirmos o conteúdo de Data Mining, vamos tratar de alguns termos complementares presentes da literatura especializada que ainda não foram vistos ao longo da nossa aula.

Alguns autores classificação mineração de dados de acordo com a forma. Essa classificação possui três categorias: Preditivo, Textual e Espacial. Veja a definição de cada um deles abaixo:

Preditivo - A data mining pode mostrar como certos atributos dos dados irão se comportar no futuro.

Textual - Processo de obtenção de informação utilizando fontes de dados textuais. Aplicações em classificação automática de textos e busca de agrupamentos.

Espacial - Processo de descoberta de padrões utilizando bancos de dados espaciais povoados por mapas.

A mineração de dados apoia o conhecimento indutivo, que descobre novas regras e padrões nos dados fornecidos. O conhecimento pode ser representado de muitas formas:

1. Quando não estruturado, pode ser representado por regras ou por lógica proposicional.
2. Em uma forma estruturada, podem ser representados por árvores de decisão, redes semânticas, redes neurais ou hierarquias de classes ou frames.

OLAP x Data Mining

O termo para processamento analítico on-line representa a característica de trabalhar os dados com operadores dimensionais. OLAP possibilita uma forma múltipla e combinada de análise.

Data Mining está mais relacionado com os processos de análise de inferência do que com a análise dimensional de dados. Representa uma forma de busca de informação baseada em algoritmos que objetivam o reconhecimento de padrões escondidos nos dados. Esses padrões não são necessariamente revelados pelas outras abordagens analíticas, como o OLAP.

Para finalizar, vamos apresentar uma última definição de data mining: “A mineração de dados é um campo interdisciplinar que reúne técnicas de **aprendizado de máquina**, **reconhecimento de padrões**, **estatísticas**, **banco de dados** e **visualização** para abordar a questão da **extração de informações** a partir de **grandes** bases de dados”.

MINERAÇÃO DE TEXTO

Considerada uma evolução da área de Recuperação de Informações (RI), a Mineração de textos (Text Mining) é um processo que utiliza técnicas de análise e extração de dados a partir de textos, frases ou apenas palavras. Envolve a aplicação de algoritmos computacionais que processam textos e identificam informações úteis e implícitas, que normalmente não poderiam ser recuperadas utilizando métodos tradicionais de consulta, pois a informação contida nestes textos não pode ser obtida de forma direta, uma vez que, em geral, estão armazenadas em formato não estruturado.

Os benefícios da mineração de textos podem se estender a qualquer domínio que utilize textos, sendo que suas principais contribuições estão relacionadas à busca de informações específicas em documentos, à análise qualitativa e quantitativa de grandes volumes de textos, e a melhor compreensão do conteúdo disponível em documentos textuais.

A ferramenta de busca do Google é um ótimo exemplo de mineração de texto. Existem vários algoritmos que classificam e ordenam os textos que aparecem como resultado das nossas consultas. Textos estes que podem estar representados das mais diversas formas, dentre elas: e-mails; arquivos em diferentes formatos (pdf, doc, txt, por exemplo); páginas Web; campos textuais em bancos de dados; textos eletrônicos digitalizados a partir de papéis.

Existem várias definições para mineração de textos. Segundo Lopes, o termo se refere ao processo de extração de padrões interessantes e não triviais, ou conhecimento a partir de

documentos em textos não estruturados. Moura descreve a mineração de textos como sendo uma área de pesquisa tecnológica cujo objetivo é a busca por padrões, tendências e regularidades em textos escritos em linguagem natural.

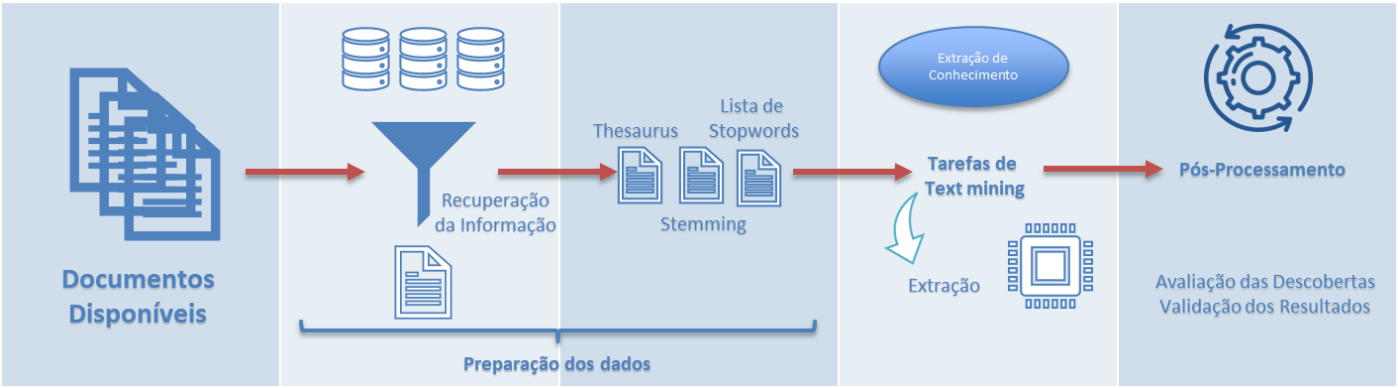


Figura 3 - Processo de mineração de texto

Já Wives afirma que a mineração de textos pode ser entendida como a aplicação de técnicas de KDD/mineração de dados sobre dados extraídos de textos. Na prática, a mineração de textos define um processo que auxilia na descoberta de conhecimento inovador a partir de documentos textuais, que pode ser utilizado em diversas áreas do conhecimento. Gostaria de falar um pouco sobre algumas ações que a ocorrem durante a preparação de dados do processo de mineração de texto.

HORA DE PRATICAR!

(Ministério da Economia – Especialista em Ciência de Dados - 2020) Julgue os seguintes itens, a respeito de big data

A mineração de textos utiliza técnicas diferentes da mineração de dados, tendo em vista que os textos representam um tipo específico de dado.

Comentários: Mineração de textos, também conhecido como mineração de dados textuais ou descoberta de conhecimento de bases de dados textuais, em geral, se refere ao processo de extração de informações de interesse e padrões não-triviais ou descoberta de conhecimento em documentos de texto não-estruturados. Pode ser visto como uma extensão da mineração de dados. A mineração de texto segue, em última instância o mesmo conjunto de etapas da mineração de dados e utiliza-se de um conjunto de técnicas comuns.

Gabarito: ERRADO.

O manuseio de arquivos texto apresenta alguns desafios. O primeiro deles envolve o **formato dos textos** com nenhuma ou pouca estruturação, o que dificulta a utilização imediata. Um outro desafio diz respeito ao tamanho dos arquivos em formato texto, comumente da ordem de milhares de palavras ou termos. Além disso, muitas dessas palavras são repetidas, expressam o mesmo significado ou possuem significado irrelevante.

As situações mencionadas acima, assim como outras, encontradas quando se lida com dados textuais, devem ser trabalhadas e resolvidas para viabilizar o uso de arquivos texto em primeira instância e em segunda aumentar a eficiência de atividades executadas a

posteriori. A preparação dos textos é a primeira etapa do processo de descoberta de conhecimento em textos.

Esta etapa envolve a seleção das bases de textos que constituirão os dados de interesse e o trabalho inicial para tentar **selecionar o núcleo que melhor expressa o conteúdo dos textos**, ou seja, toda a informação que não refletir nenhuma ideia considerada importante poderá ser desprezada. Além de promover **uma redução dimensional**, esta etapa tenta identificar similaridades em função da morfologia ou do significado dos termos, de modo a agrupar suas contribuições. Vejamos algumas ações importantes.

Uso do Thesaurus (ou dicionário)

Um **dicionário** pode ser definido como um vocabulário controlado que representa sinônimos, hierarquias e relacionamentos associativos entre termos para ajudar os leitores a encontrar a informação de que eles precisam. A pergunta é: para que eu quero montar um dicionário de palavras neste contexto?

Embora isto pareça estranho, é apenas uma questão de se entender para que serve um thesaurus. O valor do thesaurus vem justamente dos problemas inerentes à procura e indexação da linguagem natural. Usuários diferentes definem a mesma query usando termos diferentes. Para resolver este problema, um thesaurus mapeia termos variantes – sinônimos, abreviações, acrônimos, e ortografias alternativas – para um termo preferido único para cada conceito.

Para processos de indexação de documentos, o thesaurus informa que termos índices devem ser usados para descrever cada conceito. Isto reforça a consistência da indexação. Essa etapa de pré-processamento auxilia no fornecimento de um vocabulário padrão para indexação e pesquisa. O uso de um tesauro, também conhecido como uma coleção de sinônimos, tem um impacto substancial a **revocação**[1] de consulta dos sistemas de informação. Esse processo pode ser complicado porque muitas palavras possuem significados diferentes em variados contextos.

Um thesaurus pode também representar a riqueza dos relacionamentos associativos e hierárquicos. Usuários podem expressar a necessidade de informação com um nível de especificidade mais restrito ou mais amplo que o usado pelo indexador para descrever os documentos. O mapeamento de relacionamentos hierárquicos endereça este problema. Veja figura abaixo representando um tipo de relacionamento hierárquico.

[1] A revocação é definida como o número de documentos relevantes recuperados por uma pesquisa dividido pelo número total de documentos relevantes existentes.

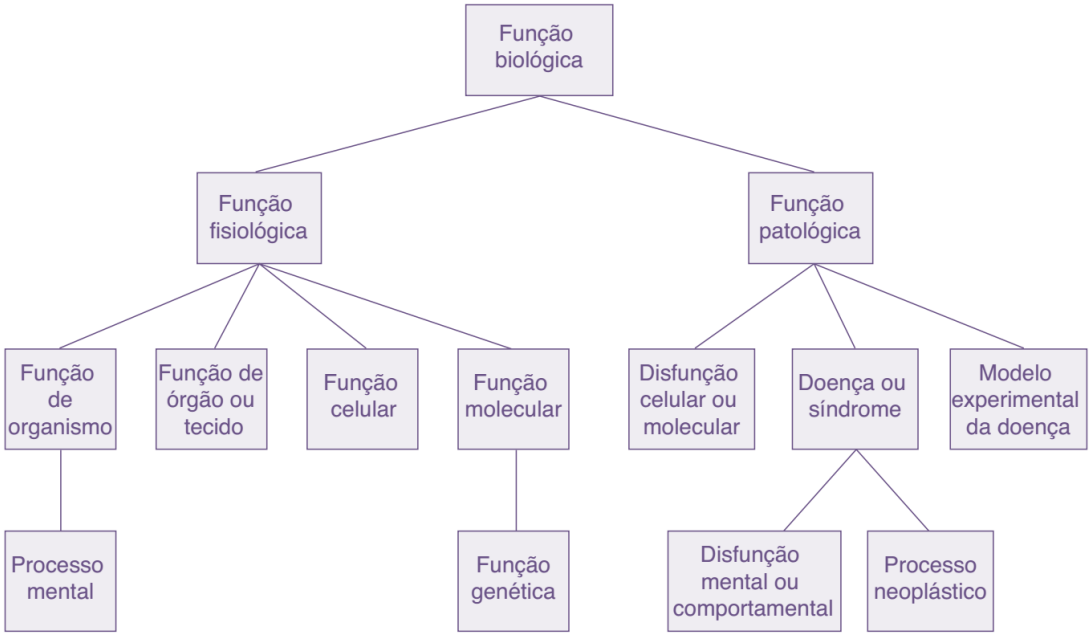


Figura 4 - Figura representando a hierarquia da função biológica.

Case Folding

Case Folding é o processo de converter todos os caracteres de um documento no mesmo tipo de letra – ou todas maiúsculas ou minúsculas. Isso tem a vantagem de acelerar comparações no processo de indexação.

Stop words

Stopwords são muito utilizadas em um idioma e desempenham um papel importante na formação de uma sentença, mas raramente contribuem para o significado dessa sentença. Palavras que se espera que ocorram em 80 por cento ou mais dos documentos em uma coleção costumam ser chamadas de ***stopwords***, e elas se tornam ***potencialmente inúteis***. Por serem muito comuns e devido à função dessas palavras, elas não contribuem muito para a relevância de um documento para uma pesquisa. Alguns exemplos (em inglês) são palavras como ***the, of, to, a, and, in, said, for, that, was, on, he, is, with, at, by e it***.

Segundo o Navathe, a remoção de stopwords de um documento **deve ser realizada antes da indexação**. Artigos, preposições, conjunções e alguns pronomes geralmente são classificados como ***stopwords***. As consultas também devem ser pré-processadas para remoção de stopword antes do processo de recuperação real. A remoção de stopwords resulta na eliminação de possíveis índices falsos, **reduzindo assim o tamanho de uma estrutura de índice em cerca de 40 por cento ou mais**.

HORA DE PRATICAR!

(Ministério da Economia – Especialista em Ciência de Dados - 2020) No que se refere à mineração de dados, julgue os itens a seguir.

O objetivo da etapa de pré-processamento é diminuir a quantidade de dados que serão analisados, por meio da aplicação de filtros e de eliminadores de palavras.

_____

Comentários: No contexto de mineração de texto, de fato, a existência de filtros e eliminadores de palavras (Stopwords, por exemplo) são usados para diminuir a quantidade de dados a serem analisadas sem prejudicar o resultado da mineração.

Gabarito: CERTO.

Stemming ou lematização (Raízes)

A **raiz ou radical** de uma palavra é definida como a palavra obtida depois de remover o sufixo e o prefixo de uma palavra original. Por exemplo, 'comput' é a palavra raiz para computador, computação e computadorizado. Esses sufixos e prefixos são muito comuns no idioma português, para dar suporte à noção de verbos, tempos e formas no plural. As raízes reduzem as diferentes formas da palavra formada por inflexão (devido a plurais e tempos) e derivação a uma raiz comum.

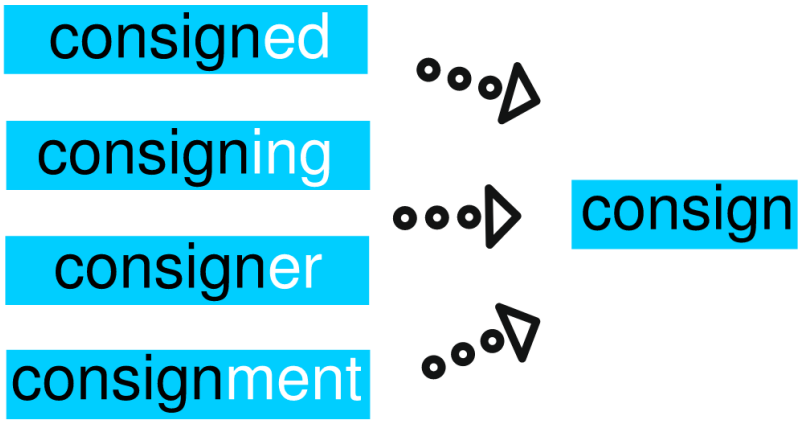


Figura 5 - Exemplo de Stemming

O processo de **stemming** é realizado considerando cada palavra isoladamente e tentando reduzi-la a sua provável palavra raiz. Isto tem a vantagem de eliminar sufixos, indicando formas verbais e/ou plurais; entretanto, algoritmos de ***stemming*** empregam linguística e são dependentes do idioma. Os algoritmos de ***stemming*** correntes não costumam usar informações do contexto para determinar o sentido correto de cada palavra, e realmente essa abordagem parece não ajudar muito.

Casos em que o contexto melhora o processo de stemming não são muito frequentes, e a maioria das palavras pode ser considerada como apresentando um significado único. Os erros resultantes de uma análise de sentido imprecisa das palavras, em geral, não compensam os ganhos que possam ser obtidos pelo aumento de precisão do processo de stemming.

Existem, porém, outros tipos de erros que devem ser observados e controlados durante a execução do stemming. Os erros mais comuns associados ao processo de stemming podem ser divididos em dois grupos:

· ***Overstemming***: acontece quando a cadeia de caracteres removida não era um sufixo, mas parte do ***stem***. Isto pode resultar na conflação de termos não relacionados.

· ***Understemming***: acontece quando um sufixo não é removido. Isto geralmente causa uma falha na conflação de palavras relacionadas

Para finalizar, vamos apresentar uma última definição de **data mining**: "A mineração de dados é um campo interdisciplinar que reúne técnicas de **aprendizado de máquina**, **reconhecimento de padrões**, **estatísticas**, **banco de dados** e **visualização** para abordar a questão da **extração de informações** a partir de **grandes** bases de dados".

(Ano: 2015 Banca: CESPE Órgão: DEPEN Prova: Agente Penitenciário Federal - Área 7) Acerca de data warehouse e data mining, julgue o item subsequente.

Os objetivos do data mining incluem identificar os tipos de relacionamentos que se estabelecem entre informações armazenadas em um grande repositório.

Comentário: Essa questão finaliza nossa parte teórica de mineração de dados. Ao analisar o texto, podemos verificar que essa é uma afirmação coerente com o assunto que vimos até aqui. Logo, a alternativa está correta.

Gabarito: C.

NOÇÕES DE APRENDIZADO DE MÁQUINA

Vamos começar essa parte da aula tentando responder a uma pergunta simples: o que é **Aprendizado de Máquina**?

Machine Learning (ML) ou aprendizado de máquina é **uma área da inteligência artificial** cujo objetivo é o desenvolvimento de **técnicas computacionais sobre o aprendizado,** bem como a construção de sistemas capazes de adquirir conhecimento de forma automática. Um sistema de aprendizado é um programa de computador que toma decisões baseado em experiências acumuladas por meio de solução bem-sucedida de problemas anteriores. É uma ferramenta poderosa para aquisição automática de conhecimento, entretanto, não existe um único algoritmo ou solução que apresente melhor desempenho para todos problemas.

Segundo *T. Michell*, o aprendizado de máquina trata do projeto e desenvolvimento de algoritmos que imitam o comportamento de aprendizagem humano, com um foco principal em aprender automaticamente a reconhecer padrões complexos e tomar decisões.

Se voltarmos algumas páginas e relembrarmos das tarefas de mineração de dados listadas no início da aula, veremos que todas elas podem ser abordadas por algum algoritmo de aprendizado de máquina. Nesta parte da aula, vamos apresentar alguns desses algoritmos sempre apresentando as tarefas que esses procuram resolver.

A aplicações atuais de ML passam por diferentes escopos. Para termos uma ideia de algumas aplicações que temos hoje em dia, vamos observar o quadro abaixo:

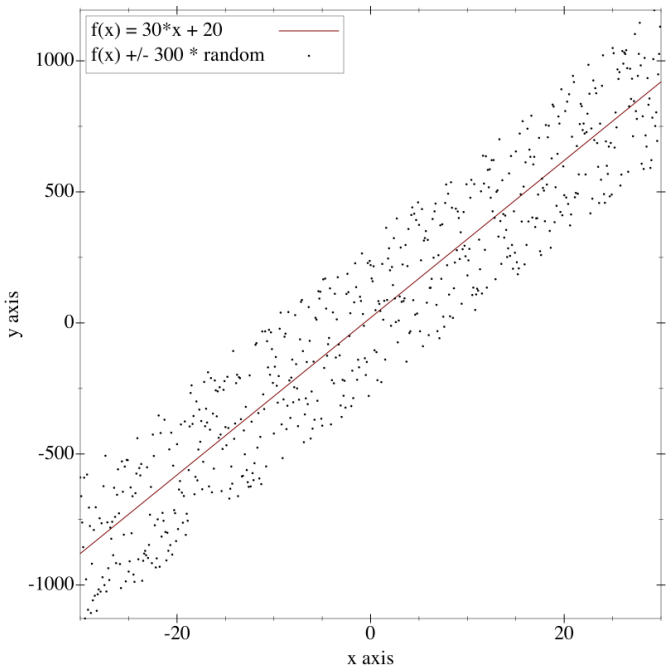
Classificação de Padrões (ex.: imagens, crédito bancário)
Processamento de Linguagem Natural (ex.: como usar notícias como fonte de informação)
Reconhecimento de Objetos (ex.: reconhecer pessoas ou objetos)
Jogar jogos (ex.: xadrez, gamão)
Previsão de variáveis relevantes (ex.: inflação, retornos de ativos financeiros)
Clustering (ex.: segmentar clientes de uma empresa)

Veja a o aprendizado de máquina trabalha como uma ciência que aborda o desenvolvimento de algoritmos e técnicas que permitem computadores/máquinas aprenderem a realizar tarefas, fazer escolhas ou prever resultados.

Duas fontes de erros devem ser analisadas quando estamos trabalhando com algoritmos de aprendizado de máquina: viés (bias) e variância. O **viés** aparece quando o algoritmo aprende um modelo incorreto, muitas vezes pode ser associado à ideia de ***underfitting,*** pois não consegue induzir um modelo que se ajusta aos dados. Já variância está relacionada ao fato de o algoritmo prestar atenção a detalhes sem importância, criando um modelo complexo e pesado, levando a um termo conhecido como ***overfitting***.

MODELOS DE APRENDIZADO DE MÁQUINA

Aprendizado de máquina é a interseção entre ciência da computação teoricamente sólida e dados ruidosos. Essencialmente, trata-se de máquinas que dão sentido aos dados, da mesma maneira que os humanos. O aprendizado de máquina é um tipo de **inteligência artificial** em que um algoritmo ou método extrairá padrões de dados. De um modo geral, há alguns problemas ou modelos associados ao aprendizado de máquina: supervisionado, não supervisionado e por reforço. Estes estão listados e descritos abaixo:

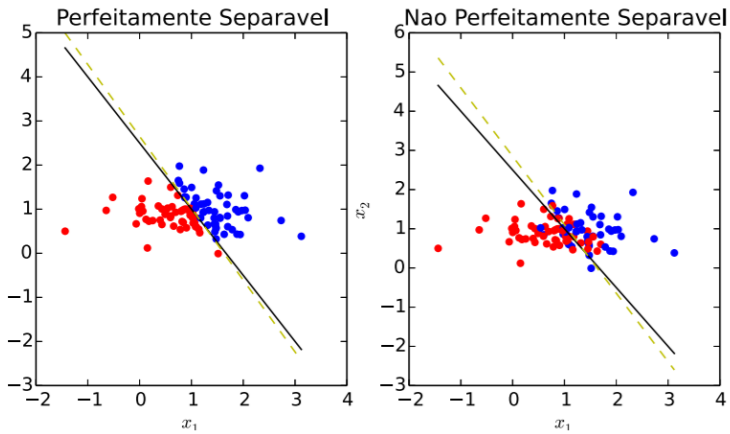


Aprendizagem Supervisionada

O aprendizado supervisionado, ou aproximação de função, procura simplesmente ajustar os dados a uma função qualquer. Por exemplo, dados os dados ruidosos (pontos fora da linha vermelha) mostrados na figura ao lado, você pode ajustar ou construir uma linha se aproxima genericamente dos dados. Uma questão interessante sobre aprendizado supervisionado é que as entradas e as saídas para um determinado conjunto de dados são conhecidas. A tarefa de classificação é um bom exemplo deste tipo de aprendizado.

Algoritmos de aprendizagem supervisionada **supõem a existência de um "professor"** que te ensina que tipo de comportamento você deve exibir em cada situação. Imagine que você deseja classificar empresas saudáveis de não saudáveis e, para fazer isso, você tem uma amostra que associa cada empresa saudável a uma série de variáveis. Então, um algoritmo de aprendizagem supervisionado tentaria usar explicitamente essa informação para, no futuro, ser hábil para separar empresas saudáveis de não saudáveis.

Agora que já sabemos o que seria o aprendizado supervisionado, vamos tentar apresentar uma técnica que trabalha este tópico. **Vamos observar o modelo PROBIT**. Na estatística, um modelo de PROBIT é um tipo de regressão em que a variável dependente pode levar apenas dois valores, por exemplo, **casado ou não**. O objetivo do modelo é estimar a probabilidade de que uma observação com características particulares apareça em uma das categorias específicas. Além disso, classificar as observações com base em suas probabilidades previstas é um tipo de modelo de classificação binária. Veja na figura abaixo um exemplo de classificação usando PROBIT, aproveitamos para introduzir o conceito de perfeitamente separável, neste caso a figura é autoexplicativa.



Quanto nem todos os pares de entrada e saída são conhecidos, o modelo pode ser descrito como aprendizado **semissupervisionado**. O aprendizado semissupervisionado é uma classe de tarefas e técnicas de aprendizado supervisionadas que também fazem uso de dados não rotulados para treinamento - normalmente, uma pequena quantidade de dados rotulados com uma grande quantidade de dados não rotulados.

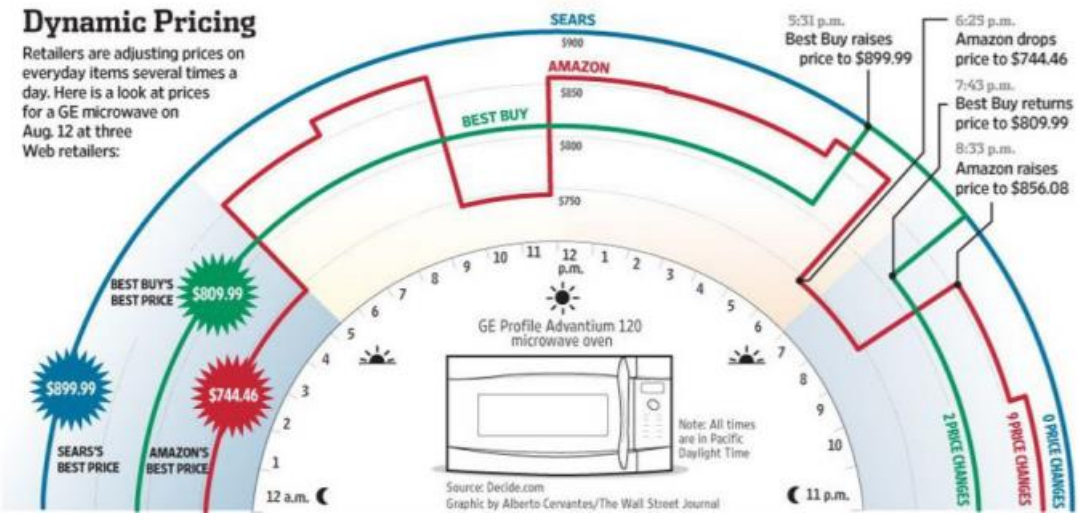
Muitos pesquisadores de aprendizado de máquina descobriram que dados não rotulados, quando usados em conjunto com uma pequena quantidade de dados rotulados, podem produzir melhorias consideráveis na precisão da aprendizagem.

Aprendizagem não supervisionada

O aprendizado não supervisionado envolve descobrir o que torna os dados especiais. Por exemplo, se recebêssemos muitos pontos de dados, poderíamos agrupá-los por semelhança, ou talvez determinar quais variáveis são melhores que outras. Nenhum rótulo é dado ao algoritmo de aprendizado, deixando-o sozinho para encontrar a estrutura em sua entrada. O aprendizado não supervisionado pode ser um objetivo em si (descobrir padrões ocultos nos dados) ou um meio para um fim.

Aprendizagem por Reforço

O aprendizado por reforço envolve descobrir como jogar um jogo de múltiplos estágios com recompensas. Os dados de treinamento (em forma de recompensas e punições) são dados apenas como feedback às ações do programa em um ambiente dinâmico, como dirigir um veículo ou jogar um jogo contra um adversário. Um exemplo interessante de aprendizado por reforço é a forma com as empresas ajustam seus preços de forma dinâmica reagindo aos preços dos concorrentes. Veja a figura abaixo:



CONCEITOS E DEFINIÇÕES

Alguns conceitos são importantes para o entendimento do assunto, em especial dos classificadores. Outros serão apresentados ao longo da aula apenas quando necessários. Vejamos no quadro abaixo:

Exemplo (caso, registro ou dado) é uma tupla de valores de atributos. Por exemplo: um paciente, dados médicos sobre uma determinada doença.

Atributo: descreve uma característica ou um aspecto de um exemplo. Por exemplo: Nominal: cor, Contínuo: peso.

Classe: atributo especial (usado no aprendizado supervisionado), também denominado de rótulo. Por exemplo: C_1, C_2, …, C_k.

Conjunto de exemplos: Um conjunto de exemplos é composto por exemplos contendo valores de atributos, bem como a classe associada.

Classificador ou Hipótese: Dado um conjunto de exemplos de treinamento, um indutor gera como saída um classificador (também denominado hipótese ou descrição de conceito) de forma que, dado um novo exemplo, ele possa predizer com a maior precisão possível sua classe.

Ruído: é comum, no mundo real, trabalhar com dados imperfeitos. Eles podem ser derivados do próprio processo que gerou os dados, do processo de aquisição de dados, do processo de transformação ou mesmo devido a classes rotuladas incorretamente (por exemplo, exemplos com os mesmos valores de atributos, mas com classes diferentes).

Missing Values (Valores Perdidos): em geral, indicados por valores fora do escopo. Tipos: desconhecidos, não registrados, irrelevantes. Razões: mau funcionamento do equipamento, mudanças na definição do experimento, incapacidade de mensuração. Obs: valores perdidos podem, de fato, significar alguma coisa. A maioria dos métodos de aprendizado não assumem isto. No entanto, este tipo de informação pode ser codificado como um valor adicional.

Modo de aprendizado:

Não incremental (batch): sempre que todo o conjunto de treinamento deva estar presente para o aprendizado.

Incremental: o indutor apenas tenta atualizar a hipótese antiga sempre que novos exemplos são adicionados ao conjunto de treinamento.

Taxa de Erro de um classificador h: Compara a classe verdadeira de cada exemplo com o rótulo atribuído pelo classificador induzido.

Acredito que você já deve ter percebido a existência de um conjunto de valores usados para treinamento do nosso modelo e outro conjunto usado para teste. Existe também um conjunto de técnicas e algoritmos para resolver o problema de aprendizado. Uma opção interessante

é combinar diferentes modelos para achar o melhor resultado. Existem alguns conceitos importantes associados a essa ideia, vejamos:

> ***Bagging***: Gera-se subamostras de seu conjunto de treino, aprenda um modelo para cada subamostra e depois combine os resultados.
>
> ***Boosting***: Usa-se pesos para combinar os modelos.
>
> ***Stack***: Outputs de vários modelos se tornam entradas de um outro modelo que escolhe a melhor forma de combiná-los.

Obs.: Hoje em dia, os melhores resultados são obtidos pela combinação de diferentes algoritmos usando os conceitos apresentados acima.

ALGORITMOS OU TÉCNICAS DE APRENDIZADO

Para cada um dos paradigmas em negrito (simbólico, estatístico, baseado em exemplos, conexista e evolutivo) apresentaremos exemplos de soluções (técnicas ou algoritmos) que envolvem aprendizado e os conceitos básicos associados a cada uma delas. Os diversos sistemas de AM possuem características que possibilitam sua classificação segundo várias dimensões: modo, paradigma, forma de aprendizado e linguagem de descrição utilizada para descrever exemplos e conhecimento**. Vamos tentar organizar essas taxonomias:**

Modos	Paradigmas	Formas	Linguagem de Descrição
• Supervisionado • Não supervisionado • Por reforço	• Simbólico • Estatístico • Baseado em exemplos • Conexista • Evolutivo	• Incremental • Não incremental	• Exemplos ou objetos • Hipóteses • Conhecimento do domínio

Com relação à forma de aprendizado, o algoritmo é não incremental quando necessita que todos os exemplos utilizados pelo algoritmo estejam simultaneamente disponíveis. Esse requerimento não é necessário para um algoritmo incremental, o qual tenta atualizar a hipótese corrente sempre que novos exemplos são adicionados ao conjunto de treinamento. Qualquer que seja o tipo de aprendizado, são necessárias linguagens para descrever exemplos, hipóteses e conhecimento do domínio. Descrever essas linguagens foge do escopo da nossa aula. Vamos seguir falando um pouco mais sobre os paradigmas.

Paradigma Simbólico - Os sistemas de aprendizado simbólico têm a característica de expressar a hipótese (conceito) induzida em uma linguagem de fácil interpretação para o usuário.

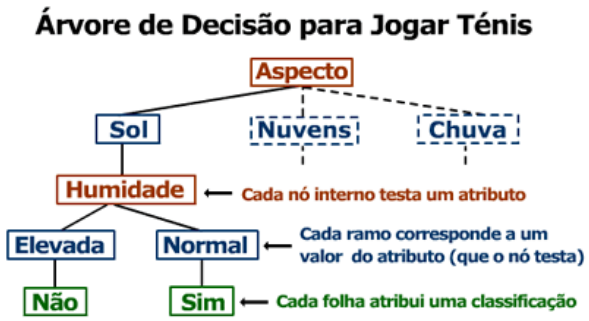
Árvores de decisão

Árvore de Decisão (AD) é um dos métodos mais consagrados em aprendizado de máquina simbólico supervisionado. Algoritmos que induzem árvores de decisão pertencem à família de algoritmos ***Top Down Induction of Decision Trees*** (TDIDT). Uma árvore de decisão é uma estrutura de dados definida recursivamente como:

• um nó folha que corresponde a uma classe; ou

• um nó de decisão que contém um teste sobre algum atributo. Para cada resultado do teste existe uma aresta para uma subárvore. Cada subárvore tem a mesma estrutura que a árvore.

Veja a figura ao lado.



Para classificar um novo exemplo, basta começar pela raiz da arvore (nó inicial da árvore), seguindo cada nó de decisão de acordo com o valor do atributo do novo exemplo até que uma folha seja alcançada. Quando uma folha é alcançada, a classificação é dada pela classe correspondente ao nó folha.

HORA DE PRATICAR!

(Ministério da Economia – Especialista em Ciência de Dados - 2020) No que se refere à mineração de dados, julgue os itens a seguir.

Estratificação é a abordagem da técnica de árvore de decisão que determina as regras para direcionar cada caso a uma categoria já existente

Comentários: Uma árvore de decisão é uma estrutura que inclui um nó raiz, ramos e nós folha. Cada nó interno denota um teste em um atributo, cada ramificação denota o resultado de um teste e cada nó folha contém um rótulo de classe. O nó superior da árvore é o nó raiz. O objetivo da técnica é andar pela árvore verificando cada um dos testes e chegar a uma folha. Essa folha vai apresentar a categoria ou rótulo do item avaliado.

Gabarito: CERTO.

Indução de regras ou Rede Semântica

A indução de uma árvore de decisão divide recursivamente os exemplos em subconjuntos menores, tentando separar cada classe das demais. Em contrapartida, a indução de regras tenta separar as classes diretamente. Nesse processo, cada nova regra cobre um subconjunto de exemplos que pertencem a uma classe específica. Basicamente, há duas formas de indução de regras: regras ordenadas e não ordenadas.

A indução de regras ordenadas trabalha de forma iterativa. Cada iteração procura por uma regra que cobre um grande número de exemplos de uma mesma classe C_i e poucos exemplos de outras classes C_j, $j \neq i$. Ao encontrar uma regra, os exemplos cobertos que pertencem à classe C_i (assim como, eventualmente, alguns exemplos de outras classes C_j,

$j \neq i$ que também são cobertos pela mesma regra) são removidos do conjunto de treinamento e a regra é adicionada no final da lista de regras. Esse processo se repete até que os exemplos de treinamento acabem ou algum critério de parada seja atingido.

Para classificar um novo exemplo, cada regra é testada em ordem — da primeira para a última — até, e se, encontrar uma (daí o nome ordenada) cujas condições sejam satisfeitas pelo novo exemplo. Já o algoritmo para induzir regras não ordenadas tem uma alteração básica que consiste em remover apenas os exemplos cobertos pela regra que são da classe C_i. Assim, diferentemente de regras ordenadas, os exemplos das classes C_j, $j \neq i$ incorretamente cobertos pela regra devem permanecer porque agora cada nova regra deve ser comparada com todos os exemplos cobertos incorretamente.

Exemplos cobertos que possuem a mesma classe C_i sendo aprendida devem ser removidos para evitar que o algoritmo encontre a mesma regra novamente. Para classificar um novo exemplo, todas as regras são testadas independentemente da ordem que foram induzidas (daí o nome de não ordenada) e todas aquelas regras que disparam são coletadas.

Veja, no exemplo abaixo, um conjunto de regras, na figura, os exemplos T_i podem ser cobertos corretamente (CC) ou cobertos incorretamente (CI).

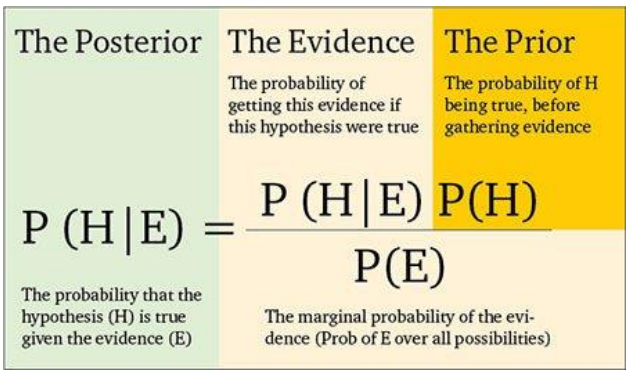
Regras ordenadas para os exemplos de viagem

	Regra	CC	CI
R_1	**if** umidade $<$ 83 **then** classe = vá	$T_1, T_3, T_7, T_8, T_9, T_{14}, T_{15}$	T_{12}, T_{13}
R_2	**else if** temperatura $\geq$ 23 **then** classe = não_vá	T_2, T_4, T_5	T_6
R_3	**else if** aparência = chuva **then** classe = vá	T_{11}	
R_4	**else** classe = não_vá	T_{10}	

Paradigma Estatístico - Os sistemas estatísticos utilizam modelos estatísticos para encontrar uma boa aproximação do conceito induzido, destacam-se entre eles os de aprendizado Bayesiano

Aprendizado Bayesiano

O pensamento Bayesiano fornece uma abordagem probabilística para aprendizagem. Está baseado na suposição de que as quantidades de interesse são reguladas por distribuições de probabilidade. A distribuição de probabilidade é uma função que descreve a probabilidade de uma variável aleatória assumir certos valores. Logo, estamos preocupados em fornecer as probabilidades para suas respostas.



Esse método permite combinar facilmente conhecimento a priori com dados de treinamento. Decisões ótimas podem ser tomadas com base nestas probabilidades conjuntamente com os dados observados. Fornece a base para algoritmos de aprendizagem que manipulam

probabilidades, bem como para outros algoritmos que não manipulam probabilidades explicitamente.

Os classificadores Bayesianos são uma família de classificadores probabilísticos simples com base na aplicação do teorema de Bayes com forte independência entre as características. Não vamos explorar os conceitos estatísticos associados ao teorema que aparece na figura ao lado.

Paradigma Baseado em Exemplos (KNN e baseado em casos)

K-Nearest Neighbours

É um dos algoritmos de classificação mais simples. Usado para classificar objetos com base em exemplos de treinamento que estão mais próximos no espaço de características. Para utilizar o KNN, é necessário:

(1) Um conjunto de exemplos de treinamento.

(2) Definir uma métrica para calcular a distância entre os exemplos de treinamento.

(3) Definir o valor de K (o número de vizinhos mais próximos que serão considerados pelo algoritmo).

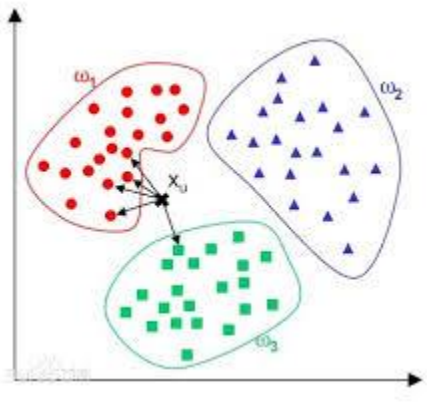
Classificar um exemplo desconhecido com o algoritmo KNN consiste em:

(1) Calcular a distância entre o exemplo desconhecido e os outros exemplos do conjunto de treinamento.

(2) Identificar os K vizinhos mais próximos.

(3) Utilizar o rótulo da classe dos vizinhos mais próximos para determinar o rótulo de classe do exemplo desconhecido (votação majoritária).

Observando a figura abaixo, o novo elemento deve ser associado a qual conjunto?

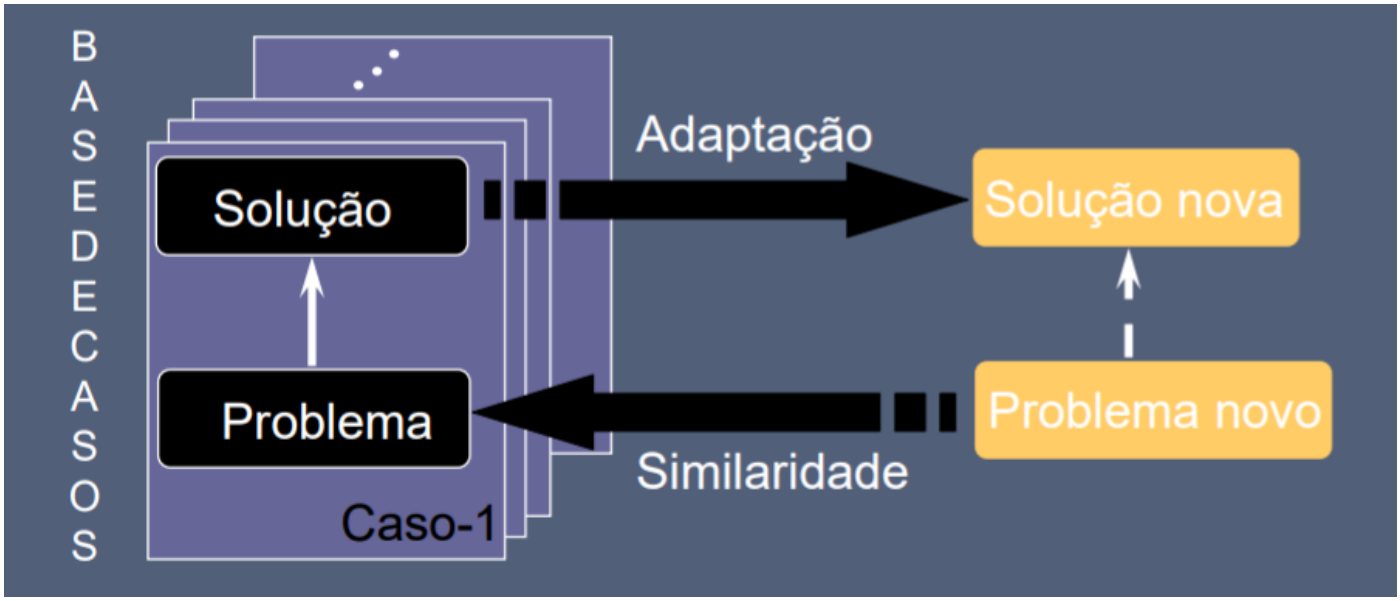


Vejam que, pela lógica apresentada acima, o novo elemento deve pertencer ao grupo vermelho.

Raciocínio baseado em casos (RBC)

Raciocínio Baseado em Casos é um enfoque para a solução de problemas e o aprendizado baseado em experiências passadas. RBC resolve problemas ao recuperar e adaptar experiências passadas – chamadas de casos – armazenadas em uma base de casos. Um novo problema é resolvido com base na adaptação de soluções de problemas similares. O

modelo resolve novos problemas pela seleção de casos com problemas semelhantes e adaptando a solução para o problema atual, conforme visto na figura abaixo.



Paradigma Conexionista (falaremos rapidamente sobre redes neurais)

Redes neurais

Uma rede neural artificial (RNA) é composta por várias unidades de processamento, cujo funcionamento é bastante simples. Essas unidades geralmente são conectadas por canais de comunicação que estão associados a determinado peso. As unidades fazem operações apenas sobre seus dados locais, que são entradas recebidas pelas suas conexões. O comportamento inteligente de uma Rede Neural Artificial vem das interações entre as unidades de processamento da rede.

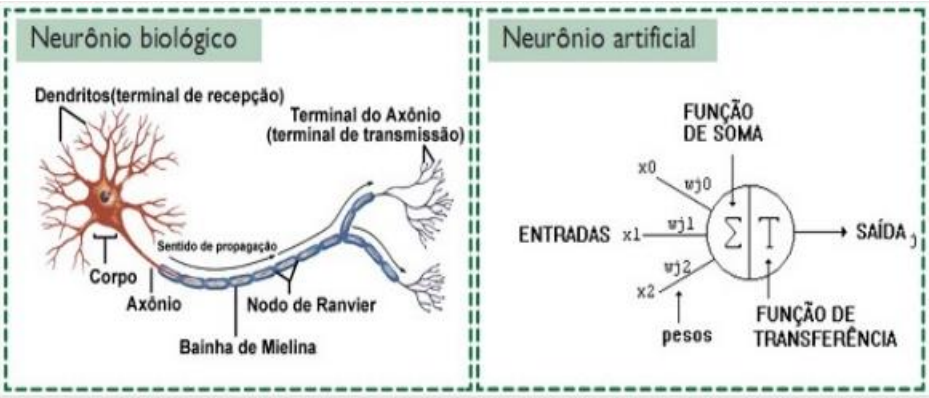
A operação de uma unidade de processamento, proposta por McCullock e Pitts em 1943, pode ser resumida da seguinte maneira:

Sinais são apresentados à entrada;

Cada sinal é multiplicado por um número, ou peso, que indica a sua influência na saída da unidade;

É feita a soma ponderada dos sinais que produz um nível de atividade;

Se este nível de atividade exceder um certo limite (threshold), a unidade produz uma determinada resposta de saída.

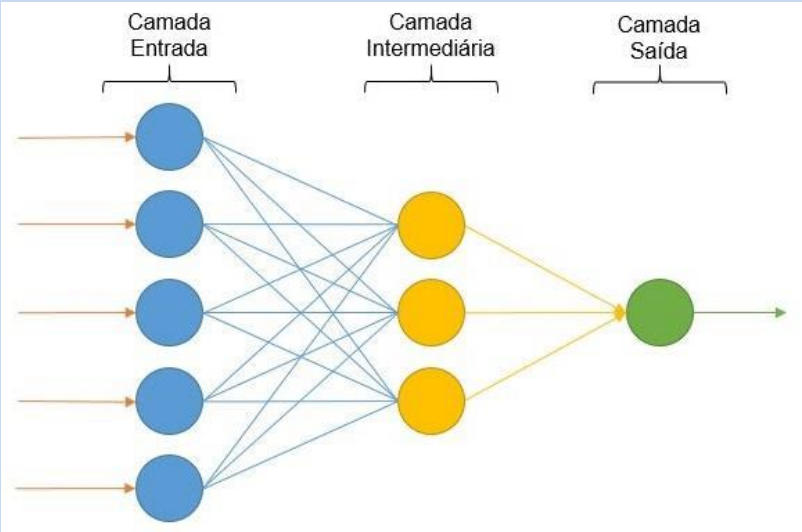


Veja o exemplo de um funcionamento da estrutura de um neurônico artificial na figura acima. A maioria dos modelos de redes neurais possui alguma regra de treinamento, em que os pesos de suas conexões são ajustados de acordo com os padrões apresentados. Em outras palavras, elas aprendem por meio de exemplos. Usualmente as camadas são classificadas em três grupos:

Camada de Entrada: onde os padrões são apresentados à rede;

Camadas Intermediárias ou Escondidas: onde é feita a maior parte do processamento, através das conexões ponderadas; podem ser consideradas como extratoras de características;

Camada de Saída: onde o resultado final é concluído e apresentado.



Representação de uma rede neural.

A maioria das RNAs contém alguma forma de 'regra de aprendizado' que modifica os pesos das conexões de acordo com os padrões de entrada que são apresentados. De certo modo, as RNAs aprendem pelo exemplo como suas contrapartes biológicas; uma criança aprende a reconhecer cães a partir de exemplos de cães.

Embora existam muitos tipos diferentes de regras de aprendizado usadas pelas redes neurais, nossa aula vai apresentar apenas a regra delta (Delta Rule). A regra delta é frequentemente utilizada pela classe mais comum de RNAs chamada *'backpropagational*

neural networks' (BPNNs). Backpropagational é uma abreviação para a propagação para trás de erro (backwards propagation of error).

Com a regra delta, como com outros tipos de retropropagação, o 'aprendizado' é um processo supervisionado que ocorre a cada ciclo (ou seja, cada vez que a rede é apresentada com um novo padrão de entrada) através de um fluxo de ativação de saídas, e a propagação de erro parar o ajuste de pesos. Simplificando, quando uma rede neural é inicialmente apresentada com um padrão, ela faz um "chute" aleatório sobre os parâmetros. Em seguida, ele vê até que ponto sua resposta foi a partir do valor atual do chute e faz um ajuste apropriado nos seus pesos de conexão (observe na figura da página anterior).

Paradigma Evolutivo (aqui nosso exemplo será os algoritmos genéticos)

Algoritmos genéticos

Os Algoritmos Genéticos (AGs) são algoritmos probabilísticos e foram inicialmente propostos pelo Professor John Holland (1975), mas somente a partir dos anos 80, é que realmente começaram a se popularizar. A ideia inicial de Holland (1975) era tentar imitar algumas etapas do processo de evolução natural das espécies incorporando-as a um algoritmo computacional.

Basicamente, o ponto de referência foi gerar, a partir de uma população de cromossomos, novos cromossomos com propriedades genéticas superiores às de seus antecedentes. Esta ideia foi então associada a soluções de um problema em que, a partir de um conjunto de soluções atuais, são geradas novas soluções superiores aos antecedentes, sob algum critério pré-estabelecido.

Algoritmos genéticos são uma classe particular de algoritmos evolutivos que usam técnicas inspiradas pela biologia evolutiva como hereditariedade, mutação, seleção natural e recombinação (ou crossing over).

Algoritmos genéticos diferem dos algoritmos tradicionais de otimização em basicamente quatro aspectos:

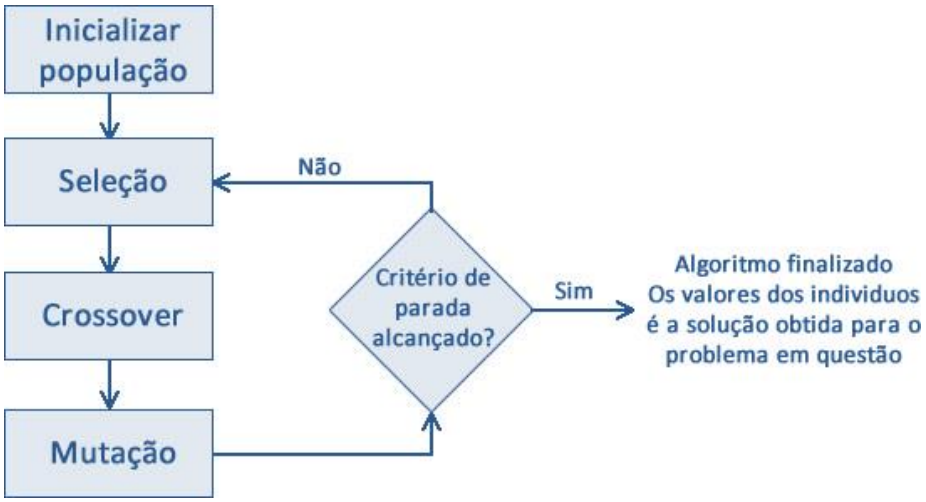
- Baseiam-se em uma codificação do conjunto das soluções possíveis, e não nos parâmetros da otimização em si;
- os resultados são apresentados como uma população de soluções e não como uma solução única;
- não necessitam de nenhum conhecimento derivado do problema, apenas de uma forma de avaliação do resultado;
- usam transições probabilísticas e não regras determinísticas.

O que você pode concluir do texto acima é que você deve olhar para o resultado e usar a probabilidade para chegar ao resultado otimizado. Gostaria que você olhasse para a figura[2] abaixo para visualizar como o fluxo de execução padrão de um algoritmo genético. Após

[2] A figura foi retirada deste artigo: http://www.computacaointeligente.com.br/algoritmos/o-algoritmo-genetico-ga/

definir a população fazemos uma seleção de um subconjunto desta que é comparada com o valor da função objetivo.

O próximo operador a ser realizado é o **crossover**. Nesta etapa é realizado o cruzamento entre indivíduos com intuito de gerar novos indivíduos. Realizado o crossover, o próximo operador é a **mutação**. Este operador também auxilia na diversidade e nada mais é do que a inserção de um ruído a alguns indivíduos da população.



QUESTÕES COMENTADAS

QUESTÕES COMENTADAS MINERAÇÃO DE DADOS

Apresentamos abaixo um conjunto de questões sobre o assunto que aprendemos nesta aula. Esperamos que elas ajudem na fixação da matéria. Qualquer dúvida, estamos às ordens!

1. (Ministério da Economia – Especialista em Ciência de Dados - 2020)

Acerca de conceitos, premissas e aplicações de big data, julgue os itens subsequentes.

No processo de agrupamento, o objeto denominado medoide é aquele que representa a mediana do grupo do conjunto.

Comentários: O algoritmo k-medoides é estendido do algoritmo k-médias para diminuir a sensibilidade para os pontos de dados outlier. No algoritmo de K-medoides, os medoides são escolhidos aleatoriamente entre o conjunto de dados. Diferentemente do k-means, onde os pontos são escolhidos aleatoriamente no espaço.

Gabarito: ERRADO.

2. Ano: 2018 Prova: Perito – Polícia Federal Banca: CESPE Assunto: Mineração de dados

Acerca de banco de dados, julgue os seguintes itens.

41 A mineração de dados se caracteriza especialmente pela busca de informações em grandes volumes de dados, tanto estruturados quanto não estruturados, alicerçados no conceito dos 4V's: volume de mineração, variedade de algoritmos, velocidade de aprendizado e veracidade dos padrões.

42 Descobrir conexões escondidas e prever tendências futuras é um dos objetivos da mineração de dados, que utiliza a estatística, a inteligência artificial e os algoritmos de aprendizagem de máquina.

Comentário: Vamos comentar cada uma das afirmações acima:

41. A afirmação mistura o conceito de mineração de dados com o conceito de Big Data. Logo, temos uma alternativa **incorreta**.

42. Questão descreve com perfeição a definição de mineração de dados, que tem por objetivos descobrir padrões úteis nos dados. Para tal, é possível utilizar ferramentas, tais como, estatísticas, inteligência artificial, aprendizagem de máquina, visualização e banco de dados. Logo, a afirmação está **correta.**

Gabarito: 41. E 42. C.

3. Ano: 2018 Banca: CESPE Assunto: Informática para Polícia Federal Cargo: Agente Conteúdo Banco de dados

Julgue os itens que se seguem, relativos a noções de mineração de dados, big data e aprendizado de máquina.

84 **Situação hipotética**: Na ação de obtenção de informações por meio de aprendizado de máquina, verificou-se que o processo que estava sendo realizado consistia em examinar as características de determinado objeto e atribuir-lhe uma ou mais classes; verificou-se também que os algoritmos utilizados eram embasados em algoritmos de aprendizagem supervisionados. **Assertiva**: Nessa situação, a ação em realização está relacionada ao processo de classificação.

Comentário: Vamos comentar a afirmação acima:

84. A situação descrita apresenta conceitos que nos levam a um modelo de **aprendizado supervisionado**. Veja que o texto fala na presença de classes, previamente definidas, que são usadas para rotular novos elementos de dados. Na assertiva, ele afirma que a situação nos leva a um processo ou tarefa de classificação. Logo, a alternativa está **correta**.

Gabarito: 84. C

4. Ano: 2018 Banca: CESPE Assunto: Informática para Polícia Federal Cargo: Agente Conteúdo Banco de dados

Julgue os itens que se seguem, relativos a noções de mineração de dados, big data e aprendizado de máquina.

86 Pode-se definir mineração de dados como o processo de identificar, em dados, padrões válidos, novos, potencialmente úteis e, ao final, compreensíveis.

Comentário: Vamos comentar a afirmação acima:

86. A questão apresenta uma definição consistente com o conceito de mineração de dados. O texto fala da descoberta de padrões úteis sobre dados. Logo, alternativa **correta**.

Gabarito: 86. C

5. Ano: 2018 Banca: CESPE Órgão: EBSERH Prova: Analista de Tecnologia da Informação

Julgue o item que se segue, a respeito de arquitetura e tecnologias de sistemas de informação.

A descoberta de novas regras e padrões em conjuntos de dados fornecidos, ou aquisição de conhecimento indutivo, é um dos objetivos de data mining.

Comentário: Vejam que essa é uma questão recente do CESPE. Ela apresenta uma definição consistente de mineração de dados. Veja que temos o aprendizado supervisionado ou indutivo, em que queremos prever um valor futuro para uma determinada variável. Ou, por outro lado, queremos descobrir padrões nos dados fornecidos. Ambos podem ser vistos como objetivos da mineração de dados. Logo, temos uma questão correta.

Gabarito: C

6. Ano: 2018 Banca: CESPE Órgão: TCM-BA Cargo: Auditor de Contas Questão: 12

Assinale a opção correta a respeito do CRISP-DM.

A CRISP-DM é uma suíte de ferramentas proprietárias que vem se tornando um padrão da indústria para mineração de dados, uma vez que fornece um plano completo e tecnologias para a realização de um projeto de mineração de dados.

B A verificação da qualidade dos dados é uma atividade da fase de entendimento dos dados.

C Durante a fase de preparação dos dados, é realizado um inventário de requisitos, suposições e restrições de recursos.

D Na fase de avaliação dos dados, são realizadas as atividades de identificar valores especiais dos dados e catalogar seu significado.

E Na fase de preparação dos dados, são realizadas as atividades de analisar o potencial de implantação de cada resultado e estimar o potencial de melhoria do processo atual.

Comentário: Vamos comentar cada uma das alternativas, lembrando que o processo do CRISP-DM é organizado em um conjunto de etapas: entendimento do negócio, seleção dos dados (*data understanding*), preparação dos dados, modelagem dos dados, avaliação do processo e execução (deployment).

A) O CRISP-DM propõe a ser uma referência que propõe uma visão geral do ciclo de vida de um projeto de mineração de dados, logo, a alternativa B está **incorreta**.

B) **Correto**! Na fase de entendimento ou seleção dos dados é traçado o perfil dos dados. Neste momento, é avaliada a qualidade dos dados.

C) A preparação dos dados está mais preocupada em trazer e ajustar os dados para execução do algoritmo de mineração. A fase de **preparação de dados** consiste na preparação dos dados que visa à **coleta, limpeza**, **transformação**, **integração** e **formatação** dos dados definidos na etapa anterior. Logo, temos mais uma alternativa **incorreta**.

D) Não existe uma fase de avaliação dos dados dentro do processo. Logo, temos mais uma alternativa **incorreta**.

E) Vejam que a alternativa A está **incorreta**, as ações descritas são associadas à fase de avaliação do processo.

Gabarito: B

7. Ano: 2018 Banca: CESPE Órgão: TCM-BA Cargo: Auditor de Contas Questão: 13

A respeito das técnicas e(ou) métodos de mineração de dados,

assinale a opção correta.

A O agrupamento (ou clustering) realiza identificação de grupos de dados que apresentam coocorrência.

B A classificação realiza o aprendizado de uma função que pode ser usada para mapear os valores associados aos dados em um ou mais valores reais.

C A regressão ou predição promove o aprendizado de uma função que pode ser usada para mapear dados em uma de várias classes discretas definidas previamente, bem como encontrar tendências que possam ser usadas para entender e explorar padrões de comportamento dos dados.

D As regras de associação identificam grupos de dados, em que os dados têm características semelhantes aos do mesmo grupo e os grupos têm características diferentes entre si.

E Os métodos de classificação supervisionada podem ser embasados em separabilidade (entropia), utilizando árvores de decisão e variantes, e em particionamento, utilizando SVM (support vector machines).

Comentário: Vamos analisar cada uma das alternativas acima:

A) A alternativa apresenta o conceito de coocorrência que está relacionado ao conceito de regra de associação, e tenta colocá-lo dentro do contexto de agrupamento. Sendo assim, temos mais uma alternativa **incorreta**.

B) A classificação baseia-se na definição de um conjunto de rótulos que permitem classificar uma nova informação em um conjunto de classes pré-definidas. Veja que essas classes não necessariamente são valores reais. Logo, temos uma alternativa incorreta.

C) A regressão não vai mapear valores em um conjunto de classes. Se você pensar que a regressão é uma função de múltiplas variáveis, a imagem desta pode ser um conjunto fechado, porém, contínuo. Logo, temos uma alternativa errada.

D) A alternativa D descreve uma característica da tarefa de clusterização. Logo, temos uma alternativa **incorreta**.

E) **Essa é a nossa resposta**. Basicamente, o SVM é um algoritmo supervisionado que tenta criar uma linha (ou uma fronteira) que melhor separa os dados. Essa linha normalmente chamamos de "**Hyperplano**". Para entender um pouco mais desse algoritmo acesse o link. Veja que ele aplica o conceito de particionamento. Agora, a ideia da classificação é construir um conceito de separabilidade entre os novos valores de entrada e as classes ou rótulos estabelecidos.

Gabarito: E.

8. CESPE - Técnico Judiciário (STJ)/Apoio Especializado/Desenvolvimento de Sistemas/2018

Julgue o item que se segue, acerca de data mining e data warehouse.
O processo de mineração de dados está intrinsecamente ligado às dimensões e a fato, tendo em vista que, para a obtenção de padrões úteis e relevantes, é necessário que esse processo seja executado dentro dos data warehouses.

Comentário: Item **ERRADO**. Houve uma mistura nas descrições de dois conceitos na questão: ***processo de mineração*** e ***data warehouse***. O processo de mineração está ligado à obtenção de padrões úteis dentro de um conjunto de dados sem classificação. Vejamos uma definição:

Mineração de dados é o processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando, assim, novos subconjuntos de dados.

Data warehouse (DW), por outro lado, possibilita a análise de grandes volumes de dados, oferecendo suporte à tarefa de tomada de decisão e planejamento. Segundo o Inmon:

Data Warehouse é uma coleção de dados orientados por assuntos, integrados, variáveis com o tempo e não voláteis, para dar suporte ao processo de tomada de decisão.

Gabarito: E

9. CESPE - Analista I (IPHAN)/Área 7/2018

Julgue o item que se segue, a respeito de tecnologias de sistemas de informação.

Na busca de padrões no data mining, é comum a utilização do aprendizado não supervisionado, em que um agente externo apresenta ao algoritmo alguns conjuntos de padrões de entrada e seus correspondentes padrões de saída, comparando-se a resposta fornecida pelo algoritmo com a resposta esperada.

Comentário: Item **ERRADO**. A descrição da assertiva refere-se a ***aprendizado supervisionado***. Segundo *Wikipedia*, no *aprendizado supervisionado,* apresenta-se ao computador exemplos de entradas e saídas previamente conhecidas cujo objetivo é aprender uma regra geral que mapeie as entradas para as saídas correspondentes "aprendidas" com os exemplos fornecidos na fase de aprendizado.

A Aprendizagem não supervisionada, por outro lado, permite-nos abordar problemas com pouca ou nenhuma ideia do que nossos resultados devem mostrar, ou seja, não conhecemos as saídas. Podemos derivar essa estrutura, agrupando os dados com base em relações entre as variáveis dos dados.

Gabarito: E

10. Ano: 2017 Banca: CESPE Órgão: TCE-PE Cargo: Auditor de Obras Públicas Questão: 119

Julgue o item que se refere a CRISP-DM (Cross Industry Standard Process for Data Mining).

[119] Durante a fase de entendimento do negócio, busca-se descrever claramente o problema, fazer a identificação dos dados e verificar se as variáveis relevantes para o projeto não são interdependentes.

Comentário: Essa para mim foi a questão mais difícil da prova. Primeiramente você teria que se lembrar das fases do CRISP, a primeira fase é o entendimento do negócio, e a segunda entendimento dos dados. Vejamos o que cada uma tem como objetivo:

Entendimento do negócio: deve determinar os objetivos de negócio, fazer uma análise da situação atual e estabelecer os objetivos da mineração de dados. Finalizando com um plano de projeto.

Entendimento dos dados: Nesta etapa vamos entender os dados baseados nos requisitos. Nesta etapa, podemos incluir uma coleta de dados, descrição, exploração e verificação da sua qualidade. Nesta etapa, temos uma característica peculiar: identificar se as variáveis do modelo são independentes umas das outras. Quando as variáveis são independentes podemos concluir que elas não possuem informações sobrepostas. Em econometria ou análise matemática, podemos pensar em variáveis que são linearmente independentes. Uma escolha cuidadosa de variáveis independentes pode fazer com que a execução dos algoritmos seja feita de forma mais eficiente.

Enfim, depois desta longa explicação teórica, podemos perceber que o examinador associou eventos de etapas diferentes dos CRISP a fase de entendimento do negócio. Logo, a alternativa está incorreta.

Gabarito: E.

11. Ano: 2017 Banca: CESPE Órgão: TCE-PE Cargo: Analista De Controle Externo Área: Auditoria De Contas Públicas Questão: 119

Em relação à análise de agrupamentos (clusterização) em mineração de dados, julgue o item seguinte.

119 O método de clustering k-means objetiva particionar 'n' observações entre 'k' grupos; cada observação pertence ao grupo mais próximo da média.

Comentário: O algoritmo de K-means é de fato um método de clusterização. Confesso que essa informação não foi vista no nosso curso. O algoritmo inicia com a escolha dos k elementos que formaram as sementes iniciais. Vamos tentar entender um pouco mais do seu funcionamento.

Escolhidas as sementes iniciais, é calculada a distância de cada elemento em relação às sementes, agrupando o elemento ao grupo que possuir a menor distância (mais similar) e recalculando o seu centroide. Quando temos mais de um elemento, imagine vários pontos em um papel, o centroide é um ponto central entre esses pontos. O processo é repetido até que todos os elementos façam parte de um dos clusters.

Depois desta etapa, fazemos ajustes nos elementos usando métodos estatísticos que tentam diminuir a dispersão dentro de cada grupo, por meio da mudança de elementos entre os grupos. O processo é interrompido quando a mudança de um elemento de um cluster para outro não gera mais ganho.

Para visualizar o algoritmo funcionando, você pode olhar na página da Wikipédia em inglês[3]. Para saber a fonte da questão, basta acessar a definição da Wikipédia em português: "Em mineração de dados, agrupamento k-means é um método de Clustering que objetiva particionar n observações dentre k grupos onde cada observação pertence ao grupo mais próximo da média. Isso resulta em uma divisão do espaço de dados em um Diagrama de Voronoi."

Desta forma, a alternativa está correta.

Gabarito: C

[3] https://en.wikipedia.org/wiki/K-means_clustering

12. Ano: 2017 Banca: CESPE Órgão: SEDF Cargo: Analista de gestão educacional – Especialidade: tecnologia da informação Questão: 119

Com relação a data mining e data warehouse, julgue os itens que se seguem.

[119] Agrupar registros em grupos, de modo que os registros em um grupo sejam semelhantes entre si e diferentes dos registros em outros grupos é uma maneira de descrever conhecimento descoberto durante processos de mineração de dados.

Comentário: Se analisarmos a descrição acima, temos uma definição da tarefa de agrupamento ou ***clustering***.

A **clusterização** é a classificação **não supervisionada de dados**, formando agrupamentos ou clusters. Ela representa uma das principais etapas do processo de análise de dados denominada análise de clusters. A análise de clusters envolve, portanto, a **organização de um conjunto de padrões** (usualmente representados na forma de vetores de atributos ou pontos em um espaço multidimensional – espaço de atributos) **em clusters**, de acordo com **alguma medida de similaridade**. De forma intuitiva, padrões pertencentes a um dado cluster devem ser mais "similares" entre si do que em relação a padrões pertencentes a outros clusters.

Vejam, portanto, que a alternativa está correta.

Gabarito: C.

13. Ano: 2016 Banca: CESPE Órgão: TRT-08 Cargo: Analista de TI - QUESTÃO 10

Acerca de data mining, assinale a opção correta.

A A fase de preparação para implementação de um projeto de data mining consiste, entre outras tarefas, em coletar os dados que serão garimpados, que devem estar exclusivamente em um data warehouse interno da empresa.

B As redes neurais são um recurso matemático/computacional usado na aplicação de técnicas estatísticas nos processos de data mining e consistem em utilizar uma massa de dados para criar e organizar regras de classificação e decisão em formato de diagrama de árvore, que vão classificar seu comportamento ou estimar resultados futuros.

C As aplicações de data mining utilizam diversas técnicas de natureza estatística, como a análise de conglomerados (cluster analysis), que tem como objetivo agrupar, em diferentes conjuntos de dados, os elementos identificados como semelhantes entre si, com base nas características analisadas.

D As séries temporais correspondem a técnicas estatísticas utilizadas no cálculo de previsão de um conjunto de informações, analisando-se seus valores ao longo de determinado período. Nesse caso, para se obter uma previsão mais precisa, devem ser descartadas eventuais sazonalidades no conjunto de informações.

E Os processos de data mining e OLAP têm os mesmos objetivos: trabalhar os dados existentes no data warehouse e realizar inferências, buscando reconhecer correlações não explícitas nos dados do data warehouse.

Comentário: Teceremos comentários sobre cada uma das alternativas acima:

A Sabemos que a mineração de dados pode acontecer sobre qualquer tipo de arquivo de dados. Lembrem-se da possibilidade de *textmining* que não tem necessidade de dados armazenados em um DW. Alternativa **errada**!

B Na alternativa B existe uma avalanche de conceitos misturados: redes neurais, que fazem parte do conjunto de assuntos relacionados à inteligência artificial; técnicas estatísticas e árvore de decisão. Cada técnica de mineração é usada com um propósito específico, por exemplo, a classificação vai permitir que você classifique novas entradas de acordo com um conjunto pré-determinado de saídas, que foram construídos em uma etapa anterior do processo. A questão peca por misturar vários conceitos.

C Criar clusters, ou seja, agrupar subconjuntos de dados de acordo com alguma semelhança entre eles. Essa é a nossa resposta.

D Uma série temporal deve considerar a sazonalidade, pela lei da oferta e demanda, se você percebe que as vendas aumentam no Natal, você pode aumentar o preço ou o estoque. O fato de desconsiderar a sazonalidade torna a questão incorreta.

E Os processos de OLAP e Data mining são diferentes em relação à complexidade e resultados esperados. OLAP é uma ferramenta de consulta em bases de dados analíticas, ele visa extrair informações por meio de queries e utilizando as operações sobre os cubos de dados, mas não aplicam algoritmos específicos neste processo. Data Mining é bem mais complexo que OLAP, ele busca padrões em grandes volumes de dados por meio de técnicas estatísticas e de algoritmos de inteligência artificial, por exemplo. Sendo assim, não é possível comparar de forma tão simplista quando a alternativa tentou fazer, por isso, a letra E está incorreta.

Gabarito: C

14. Ano: 2016 Banca: CESPE Órgão: FUNPRESP-JUD Prova: Analista - Tecnologia da Informação

Julgue o item subsecutivo, referente às tecnologias de bancos de dados.

Em DataMining, as árvores de decisão podem ser usadas com sistemas de classificação para atribuir informação de tipo.

Comentário: O algoritmo de Árvores de decisão gera uma estrutura de árvore que ajuda na classificação e na predição das amostras desconhecidas. Com base nos registros do conjunto de treinamento, uma árvore é montada e, a partir desta árvore, pode-se classificar a amostra desconhecida sem necessariamente testar todos os valores dos seus atributos. O algoritmo de classificação por árvores de decisão é considerado um algoritmo supervisionado, pois é necessário saber quais são as classes de cada registro do conjunto de treinamento.

Como o algoritmo monta uma árvore, é necessário definir quais são os elementos desta árvore. Para simplificar a explicação do algoritmo, basta pensar em uma árvore como um conjunto de nós que são conectados por ramificações. Basicamente existem três tipos de nós: o nó raiz, que inicia a árvore, os nós comuns, que dividem um determinado atributo e geram ramificações, e os nós-folha que contêm as informações de classificação do algoritmo. Já as ramificações possuem todos os valores possíveis do atributo indicado no nó para facilitar a compreensão e interpretação. Veja a figura abaixo:

A ideia do algoritmo é montar uma árvore onde cada nó indica o teste de um atributo. Os atributos escolhidos são chamados de atributos divisores ou atributos teste. A escolha de atributos é feita com base no maior ganho de informação, isto é, na qualidade de classificação do atributo. Deste modo, podemos dizer que o atributo que melhor classificar os dados deve ser escolhido como um nó da árvore. Para facilitar a compreensão, é comum colocar os valores das probabilidades de cada classe dentro do nó.

Aproveitamos a questão para introduzir maiores detalhes sobre árvore de decisão. Observamos ainda que a alternativa está plenamente correta.

Gabarito: C

15. Ano: 2016 Banca: CESPE Órgão: FUNPRESP-EXE Prova: Especialista - Tecnologia da Informação

Com relação à forma como os dados são armazenados e manipulados no desenvolvimento de aplicações, julgue o item a seguir.

Na implementação de mineração de dados (data mining), a utilização da técnica de padrões sequenciais pode ser útil para a identificação de tendências.

Comentário: Algoritmos de padrões sequenciais identificam tipos de padrões sequenciais em restrições mínimas especificadas pelo usuário. Esta técnica procura por compras ou eventos que ocorrem em uma sequência através do tempo. Por exemplo, uma loja pode descobrir que consumidores que compram TVs tendem também a comprar filmadoras de 8mm em 60% das vezes. Ou seja, pode identificar uma tendência de compras. Logo, temos uma alternativa correta!

Gabarito: C

16. Ano: 2016 Banca: CESPE Órgão: TCE-SC Cargo: Auditor de TI

Julgue os itens subsecutivos, acerca de mineração de dados.

98 Para a realização de prognósticos por meio de técnicas de mineração de dados, parte-se de uma série de valores existentes obtidos de dados históricos bem como de suposições controladas a respeito das condições futuras, para prever outros valores e situações que ocorrerão e, assim, planejar e preparar as ações organizacionais.

99 As aglomerações, tipos de informação obtidos por meio da mineração de dados, caracterizam-se por se ligarem a um único e específico evento, em torno do qual ocorrem várias ações, com produção sistêmica de informações gerenciais que apoiarão uma nova ocorrência do mesmo tipo de evento.

Comentário: A alternativa 98 trata dos aspectos de previsão relacionados à mineração de dados. Basicamente você deve escolher um algoritmo, parametrizar, utilizar dados históricos das bases de dados como entrada e tentar prever o futuro. O texto da questão está correto.

Sobre a questão 99, aglomerações ou clusters são grupos de indivíduos de uma amostra que possuem características semelhantes. Geralmente são definidos intervalos de valores, para cada intervalo temos um clusters, agrupamento ou aglomeração. Essa ideia de eventos em série relacionados a um evento inicial está relacionada a técnicas de associação. A questão, portanto, encontra-se incorreta.

Gabarito: C E

17. ANO: 2015 BANCA: CESPE ÓRGÃO: TCU PROVA: AUDITOR FEDERAL DE CONTROLE EXTERNO – ANÁLISE DE INFORMAÇÕES.

No que concerne a data mining (mineração de dados) e big data, julgue os seguintes itens.

[82] O uso prático de data mining envolve o emprego de processos, ferramentas, técnicas e métodos oriundos da matemática, da estatística e da computação, inclusive de inteligência artificial.

[83] Quem utiliza o data mining tem como objetivo descobrir, explorar ou minerar relacionamentos, padrões e vínculos significativos presentes em grandes massas documentais registradas em arquivos físicos (analógicos) e arquivos lógicos (digitais).

[84] A finalidade do uso do data mining em uma organização é subsidiar a produção de afirmações conclusivas acerca do padrão de comportamento exibido por agentes de interesse dessa organização.

[85] No ambiente organizacional, devido à grande quantidade de dados, não é recomendado o emprego de data mining para atividades ligadas a marketing.

Comentários: Vamos analisar cada uma das assertivas acima a respeito de mineração de dados.

[82] Se analisarmos qualquer definição formal de Data Mining, por exemplo: “Mineração de dados, ou data mining, é o processo de análise de conjuntos de dados que tem por objetivo a descoberta de padrões interessantes e que possam representar informações úteis”.

Segundo a teoria Data Mining é uma mistura de diferentes disciplinas: Estatística, Aprendizado de máquina e Banco de dados. Podemos, então, verificar que os termos listados na questão se enquadram perfeitamente dentro do contexto e das definições existentes. Portanto, alternativa correta.

[83] Vamos analisar mais uma definição de Data Mining para chegarmos a uma conclusão definitiva a respeito desta questão. “A mineração de dados é um campo interdisciplinar que reúne técnicas de aprendizado de máquina, reconhecimento de padrões, estatísticas, banco de dados e visualização para abordar a questão da extração de informações a partir de grandes bases de dados”. Vejam que em todas as definições as análises são feitas sobre base de dados digitais. Usar tecnologia para otimizar e ampliar o horizonte e descobrir padrões ou informações relevantes. Sendo assim, a alternativa está incorreta, pois afirma que é possível fazer análise em arquivos físicos (analógicos).

[84] A princípio, você pode ser levado a acreditar que a questão está falsa, pois Data Mining não vai produzir afirmações conclusivas. Veja, porém, que o examinador usou a palavra “subsidiar” a produção. Esse é justamente a grande característica que está descrita no detalhamento da última falácia acima. Resposta da alternativa 84 é **correta**.

[85] Observem que o examinador colocou uma afirmação incorreta. Dizer que não é possível usar DM para atividades ligadas a marketing. O caso mais clássico conhecido

é justamente o das “fraldas próximas das cervejas[4]”. Uma das maiores redes de varejo dos Estados Unidos descobriu em seu gigantesco armazém de dados que a venda de fraldas descartáveis estava associada à de cerveja. Em geral, os compradores eram homens, que saíam à noite para comprar fraldas e aproveitavam para levar algumas latinhas para casa.

Gabarito: C E C E.

18. CESPE - DEPEN - 2015 - Agente Penitenciário Federal - Tecnologia da Informação (Médio)

Acerca de datawarehouse e datamining, julgue os itens subsequentes.

[116] Os objetivos do datamining incluem identificar os tipos de relacionamentos que se estabelecem entre informações armazenadas em um grande repositório.

[117] Datamart é a denominação atribuída a um sistema de dataware que atende a áreas específicas de negócios de organizações e que representa um subconjunto lógico do datawarehouse.

[118] O datawarehouse possibilita a análise de grandes volumes de dados, que, por sua vez, permitem a realização de uma melhor análise de eventos futuros.

Comentário. Vamos, então, comentar cada uma das afirmações acima:

116. A definição dos objetivos de ***datamining*** presentes na alternativa está alinhada com as definições clássicas do conceito, vejamos uma delas: “Mineração de dados é o processo de exploração de grandes quantidades de dados com o objetivo de encontrar anomalias, padrões e correlações para suportar a tomada de decisões e proporcionar vantagens estratégicas. Usando uma ampla variedade de técnicas, você pode utilizar estas informações para aumentar as receitas, reduzir custos, melhorar o relacionamento com os clientes, reduzir riscos e muito mais”. Pelo exposto, podemos concluir que a alternativa se encontra correta.

117. O conceito de datamart está associado a um subconjunto do data warehouse e, normalmente, é orientado para uma área de negócios da empresa ou equipe específica. Considerando que os data warehouses têm seu escopo sobre toda a empresa ou organização, as informações em data marts pertencem a um único departamento. Desta forma, podemos concluir que a afirmação está correta.

118. Do ponto de vista teórico, e talvez um pouco purista, as tarefas de data mining são divididas em **descritivas** e **preditivas**. As **descritivas** caracterizam as propriedades gerais dos dados em um banco de dados, basicamente focam em achar padrões reconhecidos por seres humanos para descrever os dados.

[4] http://exame.abril.com.br/revista-exame/edicoes/633/noticias/o-que-cerveja-tem-a-ver-com-fraldas-m0053931

Já as **preditivas** realizam uma inferência sobre os dados atuais para fazer previsões sobre eles. Usam variáveis para **prever valores futuros ou desconhecidos de outras variáveis**. Vejam que você **não** faz análise de eventos futuros. Isso é semanticamente incoerente. Desmembrando a questão, temos:

1 - "O datawarehouse **possibilita** a **análise de grandes volumes de dados**." Esta frase está ok, usou a palavra **possibilita** e não disse que era **responsável**. Continuando a análise:

2 - "... que, **por sua vez**, permitem a realização de uma melhor análise de eventos futuros." Ao meu ver, esse "**que, por sua vez"** diz respeito à "**análise de grandes volumes de dados"** e não ao DW. Então, substituindo ficaria assim: "...**análise de grandes volumes de dados permitem a realização de uma melhor análise de eventos futuros"**

Veja que está estranho, a análise permite uma melhor análise ... desta forma, a questão encontra-se **errada**.

Gabarito: C C E.

19. Ano: 2015 Banca: CESPE Órgão: TJ-DFT Prova: Técnico Judiciário - Programação de Sistemas

Julgue o item a seguir, a respeito de datawarehouse e de datamining.

Em um processo de mineração, durante a etapa de preparação dos dados, são analisados os requisitos de negócio para consolidar os dados.

Comentário: Lembre-se de que o processo CRISP-DM tem uma etapa específica para **analisar os requisitos de negócio**. Em outra etapa, temos **a preparação dos dados.** Logo, a afirmação acima está incorreta.

Gabarito: E

20. Ano: 2015 Banca: CESPE Órgão: MEC Prova: Administrador de Dados

Acerca de data warehouse (DW), Business Intelligence (BI) e data mining, julgue o item que se segue.

Situação hipotética: Após o período de inscrição para o vestibular de determinada universidade pública, foram reunidas informações acerca do perfil dos candidatos, cursos inscritos e concorrências. Ademais, que, por meio das soluções de BI e DW que integram outros sistemas, foram realizadas análises para a detecção de relacionamentos sistemáticos entre as informações registradas. Assertiva: Nessa

situação, tais análises podem ser consideradas como data mining, pois agregam valor às decisões do MEC e sugerem tendências, como, por exemplo, o aumento no número de escolas privadas e a escolha de determinado curso superior.

Comentário: Observem que a afirmação está correta e de acordo com o que vimos até aqui. A mineração de dados ajuda a identificar tendências sobre os dados. Essas tarefas são conhecidas como preditivas.

Gabarito: C

21. Ano: 2015 Banca: CESPE Órgão: MEC Prova: Administrador de Banco de Dados

Julgue o item seguinte, referente a data mining.

[1] Selecionar uma amostra e determinar os conjuntos de itens frequentes dessa amostra para formar a lista de previsão de subconjunto são as principais características do algoritmo de previsão.

[2] A predição em algoritmos de data mining objetiva modelar funções sobre valores para apresentar o comportamento futuro de determinados atributos.

[3] Algoritmo genético é uma das ferramentas do data mining que utiliza mecanismos de biologia evolutiva, como hereditariedade, recombinação, seleção natural e mutação, para solucionar e agrupar problemas.

Comentário: Vamos comentar cada uma das alternativas acima.

[1] Vamos comparar o conceito de previsão com o algoritmo de amostragem para entender a diferença entre eles:

Previsão: Esta técnica tem por objetivo a avaliação de um valor de uma variável ainda não identificada, baseando-se em dados adquiridos por meio do comportamento desta variável ao longo do tempo.

Amostragem: A ideia principal do **algoritmo de amostragem** é selecionar uma pequena amostra, que caiba na memória principal do banco de dados de transações, e determinar os conjuntos de itens frequentes daquela amostra.

Veja que a alternativa 1 mistura os dois conceitos, logo, temos uma alternativa errada.

[2] Agora sim! Temos uma definição consistente de previsão, conforme apresentada no comentário acima. Logo, alternativa correta.

[3] **Algoritmos Genéricos** ou AGs são algoritmos de otimização e busca baseados nos mecanismos de seleção natural e genética. Enquanto os métodos de otimização e busca convencionais trabalham geralmente de forma sequencial, avaliando a cada instante uma possível solução, os AGs trabalham com um conjunto de possíveis soluções simultaneamente. Algoritmos Genéticos (AGs) são uma classe de procedimentos de pesquisa aleatórios capazes de realizar pesquisas adaptativas e

robustas sobre uma ampla gama de topologias de espaço de pesquisa. Modelados após o surgimento adaptativo de espécies biológicas a partir de mecanismos evolutivos e introduzidos por Holland, AGs vêm sendo aplicados com sucesso em campos diversificados como análise de imagens, escalonamentos e projetos de engenharia.

Algoritmos genéticos são uma classe particular de algoritmos evolutivos que usam técnicas inspiradas pela biologia evolutiva como hereditariedade, mutação, seleção natural e recombinação (ou crossing over). Logo, a alternativa 3 está correta.

Gabarito: E C C

EXERCÍCIOS DE MINERAÇÃO DE DADOS

1. **Ano: 2018 Prova: Perito – Polícia Federal Banca: CESPE Assunto: Mineração de dados**

Acerca de banco de dados, julgue os seguintes itens.

41 A mineração de dados se caracteriza especialmente pela busca de informações em grandes volumes de dados, tanto estruturados quanto não estruturados, alicerçados no conceito dos 4V's: volume de mineração, variedade de algoritmos, velocidade de aprendizado e veracidade dos padrões.

42 Descobrir conexões escondidas e prever tendências futuras é um dos objetivos da mineração de dados, que utiliza a estatística, a inteligência artificial e os algoritmos de aprendizagem de máquina.

2. **Ano: 2018 Banca: CESPE Assunto: Informática para Polícia Federal Cargo: Agente Conteúdo Banco de dados**

Julgue os itens que se seguem, relativos a noções de mineração de dados, big data e aprendizado de máquina.

84 **Situação hipotética**: Na ação de obtenção de informações por meio de aprendizado de máquina, verificou-se que o processo que estava sendo realizado consistia em examinar as características de determinado objeto e atribuir-lhe uma ou mais classes; verificou-se também que os algoritmos utilizados eram embasados em algoritmos de aprendizagem supervisionados. **Assertiva**: Nessa situação, a ação em realização está relacionada ao processo de classificação.

3. **Ano: 2018 Banca: CESPE Assunto: Informática para Polícia Federal Cargo: Agente Conteúdo Banco de dados**

Julgue os itens que se seguem, relativos a noções de mineração de dados, big data e aprendizado de máquina.

86 Pode-se definir mineração de dados como o processo de identificar, em dados, padrões válidos, novos, potencialmente úteis e, ao final, compreensíveis.

4. Ano: 2018 Banca: CESPE Órgão: EBSERH Prova: Analista de Tecnologia da Informação

Julgue o item que se segue, a respeito de arquitetura e tecnologias de sistemas de informação.

A descoberta de novas regras e padrões em conjuntos de dados fornecidos, ou aquisição de conhecimento indutivo, é um dos objetivos de data mining.

5. Ano: 2018 Banca: CESPE Órgão: TCM-BA Cargo: Auditor de Contas Questão: 12

Assinale a opção correta a respeito do CRISP-DM.

A CRISP-DM é uma suíte de ferramentas proprietárias que vem se tornando um padrão da indústria para mineração de dados, uma vez que fornece um plano completo e tecnologias para a realização de um projeto de mineração de dados.

B A verificação da qualidade dos dados é uma atividade da fase de entendimento dos dados.

C Durante a fase de preparação dos dados, é realizado um inventário de requisitos, suposições e restrições de recursos.

D Na fase de avaliação dos dados, são realizadas as atividades de identificar valores especiais dos dados e catalogar seu significado.

E Na fase de preparação dos dados, são realizadas as atividades de analisar o potencial de implantação de cada resultado e estimar o potencial de melhoria do processo atual.

6. Ano: 2018 Banca: CESPE Órgão: TCM-BA Cargo: Auditor de Contas Questão: 13

A respeito das técnicas e(ou) métodos de mineração de dados,

assinale a opção correta.

A O agrupamento (ou clustering) realiza identificação de grupos de dados que apresentam coocorrência.

B A classificação realiza o aprendizado de uma função que pode ser usada para mapear os valores associados aos dados em um ou mais valores reais.

C A regressão ou predição promove o aprendizado de uma função que pode ser usada para mapear dados em uma de várias classes discretas definidas previamente, bem

como encontrar tendências que possam ser usadas para entender e explorar padrões de comportamento dos dados.

D As regras de associação identificam grupos de dados, em que os dados têm características semelhantes aos do mesmo grupo e os grupos têm características diferentes entre si.

E Os métodos de classificação supervisionada podem ser embasados em separabilidade (entropia), utilizando árvores de decisão e variantes, e em particionamento, utilizando SVM (support vector machines).

7. CESPE - Técnico Judiciário (STJ)/Apoio Especializado/Desenvolvimento de Sistemas/2018

Julgue o item que se segue, acerca de data mining e data warehouse.
O processo de mineração de dados está intrinsecamente ligado às dimensões e a fato, tendo em vista que, para a obtenção de padrões úteis e relevantes, é necessário que esse processo seja executado dentro dos data warehouses.

8. CESPE - Analista I (IPHAN)/Área 7/2018

Julgue o item que se segue, a respeito de tecnologias de sistemas de informação.

Na busca de padrões no data mining, é comum a utilização do aprendizado não supervisionado, em que um agente externo apresenta ao algoritmo alguns conjuntos de padrões de entrada e seus correspondentes padrões de saída, comparando-se a resposta fornecida pelo algoritmo com a resposta esperada.

9. Ano: 2017 Banca: CESPE Órgão: TCE-PE Cargo: Auditor de Obras Públicas Questão: 119

Julgue o item que se refere a CRISP-DM (Cross Industry Standard Process for Data Mining).

[119] Durante a fase de entendimento do negócio, busca-se descrever claramente o problema, fazer a identificação dos dados e verificar se as variáveis relevantes para o projeto não são interdependentes.

10. Ano: 2017 Banca: CESPE Órgão: TCE-PE Cargo: Analista De Controle Externo Área: Auditoria De Contas Públicas Questão: 119

Em relação à análise de agrupamentos (clusterização) em mineração de dados, julgue o item seguinte.

119 O método de clustering k-means objetiva particionar ‘n’ observações entre ‘k’ grupos; cada observação pertence ao grupo mais próximo da média.

11. Ano: 2017 Banca: CESPE Órgão: SEDF Cargo: Analista de gestão educacional – Especialidade: tecnologia da informação Questão: 119

Com relação a data mining e data warehouse, julgue os itens que se seguem.

[119] Agrupar registros em grupos, de modo que os registros em um grupo sejam semelhantes entre si e diferentes dos registros em outros grupos é uma maneira de descrever conhecimento descoberto durante processos de mineração de dados.

12. Ano: 2016 Banca: CESPE Órgão: TRT-08 Cargo: Analista de TI - QUESTÃO 10

Acerca de data mining, assinale a opção correta.

A A fase de preparação para implementação de um projeto de data mining consiste, entre outras tarefas, em coletar os dados que serão garimpados, que devem estar exclusivamente em um data warehouse interno da empresa.

B As redes neurais são um recurso matemático/computacional usado na aplicação de técnicas estatísticas nos processos de data mining e consistem em utilizar uma massa de dados para criar e organizar regras de classificação e decisão em formato de diagrama de árvore, que vão classificar seu comportamento ou estimar resultados futuros.

C As aplicações de data mining utilizam diversas técnicas de natureza estatística, como a análise de conglomerados (cluster analysis), que tem como objetivo agrupar, em diferentes conjuntos de dados, os elementos identificados como semelhantes entre si, com base nas características analisadas.

D As séries temporais correspondem a técnicas estatísticas utilizadas no cálculo de previsão de um conjunto de informações, analisando-se seus valores ao longo de determinado período. Nesse caso, para se obter uma previsão mais precisa, devem ser descartadas eventuais sazonalidades no conjunto de informações.

E Os processos de data mining e OLAP têm os mesmos objetivos: trabalhar os dados existentes no data warehouse e realizar inferências, buscando reconhecer correlações não explícitas nos dados do data warehouse.

13. Ano: 2016 Banca: CESPE Órgão: FUNPRESP-JUD Prova: Analista - Tecnologia da Informação

Julgue o item subsecutivo, referente às tecnologias de bancos de dados.

Em DataMining, as árvores de decisão podem ser usadas com sistemas de classificação para atribuir informação de tipo.

14. Ano: 2016 Banca: CESPE Órgão: FUNPRESP-EXE Prova: Especialista - Tecnologia da Informação

Com relação à forma como os dados são armazenados e manipulados no desenvolvimento de aplicações, julgue o item a seguir.

Na implementação de mineração de dados (data mining), a utilização da técnica de padrões sequenciais pode ser útil para a identificação de tendências.

15. Ano: 2016 Banca: CESPE Órgão: TCE-SC Cargo: Auditor de TI

Julgue os itens subsecutivos, acerca de mineração de dados.

98 Para a realização de prognósticos por meio de técnicas de mineração de dados, parte-se de uma série de valores existentes obtidos de dados históricos bem como de suposições controladas a respeito das condições futuras, para prever outros valores e situações que ocorrerão e, assim, planejar e preparar as ações organizacionais.

99 As aglomerações, tipos de informação obtidos por meio da mineração de dados, caracterizam-se por se ligarem a um único e específico evento, em torno do qual ocorrem várias ações, com produção sistêmica de informações gerenciais que apoiarão uma nova ocorrência do mesmo tipo de evento.

16. ANO: 2015 BANCA: CESPE ÓRGÃO: TCU PROVA: AUDITOR FEDERAL DE CONTROLE EXTERNO – ANÁLISE DE INFORMAÇÕES.

No que concerne a data mining (mineração de dados) e big data, julgue os seguintes itens.

[82] O uso prático de data mining envolve o emprego de processos, ferramentas, técnicas e métodos oriundos da matemática, da estatística e da computação, inclusive de inteligência artificial.

[83] Quem utiliza o data mining tem como objetivo descobrir, explorar ou minerar relacionamentos, padrões e vínculos significativos presentes em grandes massas documentais registradas em arquivos físicos (analógicos) e arquivos lógicos (digitais).

[84] A finalidade do uso do data mining em uma organização é subsidiar a produção de afirmações conclusivas acerca do padrão de comportamento exibido por agentes de interesse dessa organização.

[85] No ambiente organizacional, devido à grande quantidade de dados, não é recomendado o emprego de data mining para atividades ligadas a marketing.

17. CESPE - DEPEN - 2015 - Agente Penitenciário Federal - Tecnologia da Informação (Médio)

Acerca de datawarehouse e datamining, julgue os itens subsequentes.

[116] Os objetivos do datamining incluem identificar os tipos de relacionamentos que se estabelecem entre informações armazenadas em um grande repositório.

[117] Datamart é a denominação atribuída a um sistema de dataware que atende a áreas específicas de negócios de organizações e que representa um subconjunto lógico do datawarehouse.

[118] O datawarehouse possibilita a análise de grandes volumes de dados, que, por sua vez, permitem a realização de uma melhor análise de eventos futuros.

18. Ano: 2015 Banca: CESPE Órgão: TJ-DFT Prova: Técnico Judiciário - Programação de Sistemas

Julgue o item a seguir, a respeito de datawarehouse e de datamining.

Em um processo de mineração, durante a etapa de preparação dos dados, são analisados os requisitos de negócio para consolidar os dados.

19. Ano: 2015 Banca: CESPE Órgão: MEC Prova: Administrador de Dados

Acerca de data warehouse (DW), Business Intelligence (BI) e data mining, julgue o item que se segue.

Situação hipotética: Após o período de inscrição para o vestibular de determinada universidade pública, foram reunidas informações acerca do perfil dos candidatos, cursos inscritos e concorrências. Ademais, que, por meio das soluções de BI e DW que integram outros sistemas, foram realizadas análises para a detecção de relacionamentos sistemáticos entre as informações registradas. Assertiva: Nessa situação, tais análises podem ser consideradas como data mining, pois agregam valor às decisões do MEC e sugerem tendências, como, por exemplo, o aumento no número de escolas privadas e a escolha de determinado curso superior.

20. Ano: 2015 Banca: CESPE Órgão: MEC Prova: Administrador de Banco de Dados

Julgue o item seguinte, referente a data mining.

[1] Selecionar uma amostra e determinar os conjuntos de itens frequentes dessa amostra para formar a lista de previsão de subconjunto são as principais características do algoritmo de previsão.

[2] A predição em algoritmos de data mining objetiva modelar funções sobre valores para apresentar o comportamento futuro de determinados atributos.

[3] Algoritmo genético é uma das ferramentas do data mining que utiliza mecanismos de biologia evolutiva, como hereditariedade, recombinação, seleção natural e mutação, para solucionar e agrupar problemas.

GABARITO QUESTÕES

1. E C
2. C
3. C
4. C
5. B
6. D
7. E
8. E
9. E
10. C
11. C
12. C
13. C
14. C
15. C E
16. C E C E
17. C C E
18. E
19. C
20. E C C

QUESTÕES COMENTADAS MINERAÇÃO DE DADOS (OUTRAS BANCAS)

1. Ano: 2018 Banca: FCC Cargo: Auditoria e Fiscalização Questão: 71.

Um Auditor da Receita Estadual pretende descobrir, após denúncia, elementos que possam caracterizar e fundamentar a possível existência de fraudes, tipificadas como sonegação tributária, que vêm ocorrendo sistematicamente na arrecadação do ICMS. A denúncia é que, frequentemente, caminhões das empresas Org1, Org2 e Org3 não são adequadamente fiscalizados nos postos de fronteiras. Inobservâncias de procedimentos podem ser avaliadas pelo curto período de permanência dos caminhões dessas empresas na operação de pesagem, em relação ao período médio registrado para demais caminhões.

Para caracterizar e fundamentar a existência de possíveis fraudes, o Auditor deverá coletar os registros diários dos postos por, pelo menos, 1 ano e elaborar demonstrativos para análises mensais, trimestrais e anuais.

71. A aplicação de técnicas de mineração de dados (data mining) pode ser de grande valia para o Auditor. No caso das pesagens, por exemplo, uma ação típica de mining, que é passível de ser tomada com o auxílio de instrumentos preditivos, é

(A) realizar uma abordagem surpresa em determinado posto, com probabilidade significativa de constatar ocorrência fraudulenta.

(B) reportar ao escalão superior as características gerais das pesagens e permanências de todos os caminhões, nos cinco maiores postos do Estado, no mês que antecede a data de análise.

(C) quantificar as ocorrências de possíveis pesagens fraudulentas ocorridas durante todo o trimestre que antecede a data da análise, em alguns postos selecionados, mediante parâmetros comparativos preestabelecidos.

(D) analisar o percentual de ocorrências das menores permanências de caminhões nos postos, no último ano, em relação ao movimento total.

(E) relacionar os postos onde ocorreram, nos últimos seis meses, as menores permanências das empresas suspeitas e informar o escalão superior para a tomada de decisão.

Comentário: O detalhe desta questão é a existência de uma análise preditiva. Toda análise preditiva carrega consigo **um grau de incerteza**. Logo, observamos na alternativa “A” a palavra-chave **probabilidade**. Veja que estamos tentando determinar a hora certa para uma abordagem surpresa. As demais alternativas apresentam

análises descritivas, algumas podem inclusive serem feitas sem a ajuda de mineração de dados, utilizando, por exemplo, ferramentas OLAP.

Gabarito: A.

2. FCC - Analista de Gestão (SABESP)/Publicidade e Propaganda/2018

O conceito de Data Mining descreve

a) o uso de teorias, métodos, processos e tecnologias para organizar uma grande quantidade de dados brutos para identificar padrões de comportamentos em determinados públicos.

b) o conjunto de métodos, tecnologias e estratégias para atração voluntária de visitantes, buscando a conversão consistente de leads em clientes (realização de compra).

c) as atividades coordenadas de modo sistemático por uma determinada organização para relacionamento com os seus distintos públicos, bem como com outras organizações, sejam públicas, privadas ou não governamentais.

d) o conjunto de tarefas e processos, organizados e sistematizados, normalmente como uso de uma plataforma tecnológica (hardware e software, ou até mesmo em cloud computing) para a gestão do relacionamento com clientes.

e) o trabalho de produzir levantamento sobre os hábitos de consumo de mídia de um determinado público, identificando horários, tempo gasto etc., associando ao perfil socioeconômico, potencial de consumo, persuasão etc.

Comentário: A questão apresenta um conceito de mineração de dados na alternativa A. Vejamos mais uma definição: mineração de dados refere-se à mineração ou à descoberta de novas informações em termos de **padrões ou regras** com base em **grandes quantidades de dados**. Para ser útil na prática, a mineração de dados precisa ser executada de modo eficiente em grandes arquivos e bancos de dados.

Vejamos o que as demais alternativas descrevem:

b) **Errada**. A descrição está associada à **marketing digital**.

c) **Errada**. Trata do gerenciamento das partes interessadas.

d) **Errada**. Descreve o gerenciamento do relacionamento com o cliente feito em um CRM.

e) **Errada**. Essa descrição está associada à descoberta de padrões de consumo, apenas uma das possíveis tarefas de mineração.

Gabarito: A.

3. FCC - Analista em Gestão (DPE AM)/Especializado em Tecnologia da Informação de Defensoria/Analista de Banco de Dados/2018

Dentre os algoritmos utilizados em data mining, há um algoritmo que visa o estabelecimento de categorias, a partir do conjunto de dados, bem como a distribuição dos dados nas categorias estabelecidas. Essa descrição corresponde aos algoritmos de

a) classificação.

b) sumarização.

c) visualização.

d) evolução.

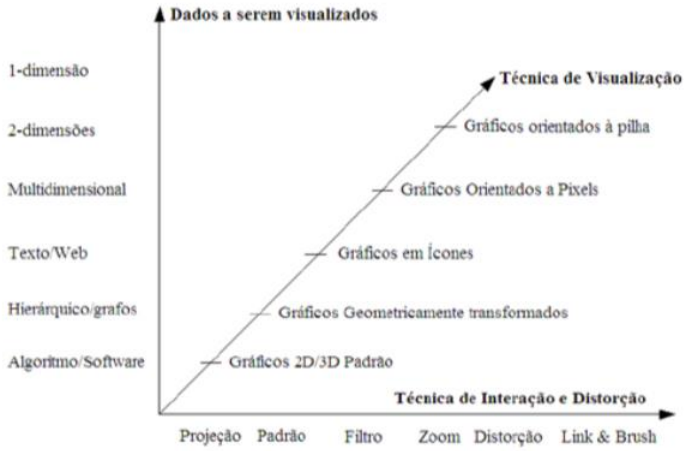
e) detecção de desvios.

Comentário: O comando da questão pede claramente um algoritmo que tem por objetivo o estabelecimento de **CATEGORIAS ou CLASSES** a partir de um conjunto de dados e a distribuição dos dados nessas categorias. Analisando as opções:

a) classificação. Visa identificar a qual classe um determinado registro pertence. O algoritmo analisa o conjunto de registros fornecidos, com cada registro já contendo a **indicação à qual classe pertence (ou seja, caracteriza),** a fim de 'aprender' como classificar um novo registro **(ou seja, aprendizado supervisionado). Essa é a nossa resposta.**

b) sumarização. Sumarização é muito usado quando há a **visualização** dos dados, como, por exemplo, somar dados de um atributo de uma tabela para realizar análises. Um dos principais objetivos do data mining é mostrar ao usuário o resultado dos dados de uma forma fácil de compreender. Com a sumarização usamos outras técnicas para melhor compreensão dos dados. Por exemplo, uma **agregação** envolve aplicações de táticas de **sumarização** de dados para reduzir o volume e agilizar processos. Logo, não caracteriza os dados, mas pode ser usado em conjunto, alternativa **ERRADA.**

c) visualização. O principal objetivo da visualização é a compreensão visual do usuário final, e não a caracterização, logo, alternativa **ERRADA.**



d) evolução. Evolução é uma técnica que descreve e estuda a regularidade de modelos ou tendências para atributos que o comportamento muda ao longo do tempo. Não há uma categorização dos dados, logo, alternativa **ERRADA.**

e) detecção de desvios. Detecção de desvios ou análise de outliers tem como objetivo **encontrar dados que não obedecem ao comportamento** ou ao modelo dos dados. Ou seja, identificação de dados que deveriam seguir um padrão esperado, mas não o fazem. Temos como exemplo um IDS, que detecta anomalias na rede para detectar intrusão de sistemas. Não há uma categorização dos dados não supervisionada, logo, alternativa **ERRADA.**

Gabarito: A

4. FCC - Analista (DPE RS)/Tecnologia da Informação/Banco de Dados/2017

Uma das técnicas bastante utilizadas em sistemas de apoio à decisão é o Data Mining, que se constitui em uma técnica

a) para a exploração e análise de dados, visando descobrir padrões e regras, a princípio ocultos, importantes à aplicação.

b) para se realizar a criptografia inteligente de dados, objetivando a proteção da informação.

c) que visa sua distribuição e replicação em um cluster de servidores, visando aprimorar a disponibilidade de dados.

d) de compactação de dados, normalmente bastante eficiente, permitindo grande desempenho no armazenamento de dados.

e) de transmissão e recepção de dados que permite a comunicação entre servidores, em tempo real.

Comentário: Uma questão direta ao ponto. Observe uma interessante descrição do conceito de data mining mostrado abaixo:

Data mining é uma expressão inglesa ligada à informática cuja tradução é mineração de dados. Consiste em uma funcionalidade que agrega e organiza dados, encontrando neles padrões, associações, mudanças e anomalias relevantes.

As demais alternativas não fazem sentido dentro do contexto de mineração, vejamos termos chaves em cada uma delas que invalidam qualquer possibilidade destas serem uma alternativa correta:

b) Criptografia.

c) Replicação

d) Compactação.

e) Transmissão e recepção de dados.

Gabarito: A

5. ANO: 2012 BANCA: FCC ÓRGÃO: TST PROVA: ANALISTA JUDICIÁRIO - ANALISTA DE SISTEMAS

Leia as afirmações a seguir:

I. Um Data Warehouse é um repositório de dados atuais e históricos de uma organização que possibilita a análise de grande volume de dados para suportar a tomada de decisões estratégicas, possuindo registros permanentes.

II. O processo de Data Mining, ou mineração de dados, tem por objetivo localizar possíveis informações em um banco de dados através de comparações com dados informados pelo usuário e registros de tabelas.

III. Um ERP, ou Sistema Integrado de Gestão Empresarial, é conhecido por integrar os dados de diferentes departamentos de uma organização, aumentando o uso de interfaces manuais nos processos.

IV. As ferramentas OLAP (On-line Analytical Processing) são capazes de analisar grandes volumes de dados, fornecendo diferentes perspectivas de visão e auxiliando usuários na sintetização de informações.

Está correto o que se afirma APENAS em

A I e II.

B II e III.

C I, III e IV.

D I, II e III.

E I e IV.

Comentário: As alternativas I e IV estão corretas. Vejamos o que está errado nas demais alternativas:

Na alternativa II temos: "O processo de *Data Mining*, ou mineração de dados, tem por objetivo localizar possíveis informações em um banco de dados por meio de comparações com dados informados pelo usuário e registros de tabelas". Já vimos a definição de data mining em outra questão, vamos aproveitar para conhecer outra forma de descrever o conceito, **mineração de dados** é o processo de descobrir relacionamentos novos, padrões e tendências por meio da análise intensiva de grandes dados históricos, utilizando inteligência artificial e técnicas estatísticas e matemáticas.

A alternativa III diz "Um ERP, ou Sistema Integrado de Gestão Empresarial, é conhecido por integrar os dados de diferentes departamentos de uma organização, aumentando o uso de interfaces manuais nos processos". Vejam que salientamos o que estava incorreto, **o ERP elimina o uso de interfaces manuais.**

Gabarito: E

1. CESGRANRIO - Analista de Sistemas Júnior (TRANSPETRO)/SAP/2018

Classificação é uma importante tarefa utilizada na etapa de mineração de dados, que tem como uma de suas características básicas

a) construir seus modelos de enquadramento, a partir de um conjunto de dados contínuos.

b) poder ser implementada por algoritmos estáveis e de significativa eficácia, tais como C4.5, classificadores bayesianos ou K-Prototypes.

c) ser um método de aprendizado de máquina não supervisionado, observando o teorema NFL – No FreeLunch.

d) ter a sua eficácia avaliada por uma métrica denominada suporte, que indica quantas vezes um item de dado foi corretamente classificado.

e) ter como seu primeiro processo o aprendizado de uma função de mapeamento y = f(X), que associa uma ocorrência de dados X em uma classe y.

Comentário: Vamos comentar cada uma das questões acima:

a) **Errada**. Supondo que com conceito de modelo de enquadramento é possível decidir, a partir de um conjunto de dados de entrada discretos, um valor ao qual esses dados de entrada são associados. Para tal, você pode usar as tarefas de classificação e regressão. Uma maneira de decidir entre as tarefas é pensar na variável de destino que você está tentando prever. Se a variável é um **número contínuo**, como o preço de um item, então você deve usar um **modelo de regressão**. Alternativamente, **se a variável é o rótulo** para diferentes **categorias de itens**, então você deve usar um modelo de **classificação**.

b) **Errado**. A alternativa fala dos métodos de implementação e não das características básicas da tarefa de classificação. Sobre os métodos descritos na questão:

O C4.5 é um dos algoritmos clássicos da tarefa de Classificação.

Goldschmidt, Ronaldo; Passos, Emmanuel. Data Mining (Locais do Kindle 4387-4388). Elsevier Editora Ltda.. Edição do Kindle.

Diversos algoritmos de Mineração de Dados são fundamentados em princípios e teorias da Estatística. O **Classificador Bayesiano Ingênuo (CBI)**, por exemplo, baseia-se no Teorema de Bayes, estando relacionado com o cálculo de probabilidades condicionais. Conforme o próprio nome sugere, este método é aplicável à tarefa de Classificação.

O **método k-Prototypes** é a integração dos métodos k-Means e k-Modes. Este método pode ser aplicado a conjuntos de dados que contenham tanto atributos numéricos quanto atributos categóricos.

c) **Errada**. É importante lembrar que a classificação é uma tarefa de aprendizados supervisionados, em que os dados e os rótulos usados para construção do modelo são conhecidos antecipadamente. Vamos aproveitar a questão para falar do almoço grátis que não existe! 😊

Segundo o teorema NFL (No Free Lunch Theorem), **não existe um algoritmo de aprendizado** que **seja superior** a todos os demais quando considerados **todos os problemas de Classificação possíveis**. Isto significa que, a cada nova aplicação envolvendo a tarefa de Classificação, os algoritmos disponíveis devem ser experimentados a fim de identificar aqueles que obtêm melhor desempenho.

d) **Errada**. Suporte é uma métrica associada ao algoritmo de regra de associação. Quando estamos tratando de classificadores, uma **medida de desempenho** muito utilizada na avaliação de classificadores é a **acurácia (total)**, também denominada **taxa de acerto do classificador**.

e) **Certo!** Sobre Classificação, temos as seguintes características:

- Busca por uma função que permita associar corretamente cada registro X(i) de um banco de dados a um único rótulo categórico Y(j), chamado de **classe**.
- Cada algoritmo possui um “bias” indutivo que direciona o processo de seleção dos classificadores, o qual influencia na seleção de hipóteses.
- **Completude**: capacidade de um classificador em classificar todos os exemplos da base de dados.
- **Consistência**: capacidade de um classificador em classificar corretamente os exemplos da base de dados.

Assim, podemos encontrar nossa resposta na alternativa E.

Gabarito: E

2. CESGRANRIO - Técnico Científico (BASA)/Tecnologia da Informação/2018

As ferramentas e técnicas de mineração de dados (data mining) têm por objetivo

a) preparar dados para serem utilizados em um “data warehouse” (DW).

b) permitir a navegação multidimensional em um DW.

c) projetar, de forma eficiente, o registro de dados transacionais.

d) buscar a classificação e o agrupamento (clusterização) de dados, bem como identificar padrões.

e) otimizar o desempenho de um gerenciador de banco de dados.

Comentário: Vamos comentar cada uma das alternativas acima:

a) **Errado**. A ferramenta de ETL – Extração, transformação e carga – que é responsável por preparar os dados armazenados em um DW.

b) **Errado**. A ferramenta OLAP que permite a navegação entre as diversas dimensões.

c) **Errado**. Os registros transacionais são projetos pela modelagem de dados relacional e orienta a construção de bases de dados OLTP – On-line Transaction Processing.

d) **Certa!** Um dos objetivos da mineração de dados é encontrar padrões úteis sobre grandes bases de dados utilizando uma das tarefas de mineração. Dentre essas tarefes podemos listar a classificação e a clusterização.

e) **Errado**. Otimização de desempenho ou Tuning é uma ação desempenhada pelo administrador de banco de dados que procura ajustar os parâmetros do banco de dados para melhorar a performance em termos de tempo de resposta e vazão dos dados.

Gabarito: D

3. CESGRANRIO - Analista (PETROBRAS)/Sistema Júnior/2018

Considere o conjunto de dados a seguir, obtido a partir de uma base de dados de transações ocorridas em uma padaria, em uma determinada faixa de tempo. Nesse conjunto, indica-se a ocorrência de um determinado produto em cada transação.

id transação	leite	café	pão	manteiga
1	não	sim	sim	sim
2	sim	não	sim	sim
3	não	sim	sim	sim
4	sim	sim	sim	sim
5	não	não	não	não
6	não	não	não	sim
7	não	não	sim	não
8	não	não	não	não
9	não	não	não	não
10	não	não	não	não

Em mineração de dados, a partir desse conjunto de dados, podem ser geradas regras de associação, as quais buscam conjuntos de itens frequentes que guardam entre si uma relação de causa e efeito. Essas regras são da forma X→Y, onde X é o antecedente da regra, e Y, seu consequente. A menção X,Y indica referência aos dois itens frequentes X e Y. Dois indicadores utilizados para averiguar a eficácia de regras de associação são o suporte e a confiança.

Qual das regras a seguir possui confiança mínima de 80%, dado um suporte mínimo de 30%?

a) café, leite → pão

b) café, pão → manteiga

c) leite, pão → café

d) manteiga → café, pão

e) manteiga, pão → café

Comentário: Para resolver essa questão você precisa avaliar o suporte e a confiança. Sugiro que você comece pelo suporte. Para calcular o suporte, basta observar entre as linhas ou transações da tabela, em quais delas os itens antecedentes e subsequentes aparecem e dividir pelo número total de linhas. Por exemplo:

café, pão → manteiga

Essa regra aparece nas linhas 1, 3 e 4. Como temos 10 linhas, podemos calcular o suporte simplesmente dividindo 3/10 = 30%.

Já a confiança é calculada da seguinte forma: Considere a quantidade de linhas em que o antecedente aparece. No caso em análise, o par (café, pão) aparece apenas nas linhas 1, 3 e 4, ou seja, em 3 das transações (A). Agora dividimos esse valor pela quantidade de linhas onde a regra se verifica. Já vimos que a regra café, pão → manteiga, verifica-se nas linhas 1, 3 e 4 (B). Assim, para calcular a confiança, dividimos B por A. Neste caso, temos uma confiança de 100%.

Para essa regra, temos o suporte de 30% e a confiança de 100%. Logo, essa é a nossa resposta. Aproveite a questão para calcular o suporte e a confiança para as demais alternativas.

Gabarito: B

4. CESGRANRIO - Analista de Sistemas Júnior (TRANSPETRO)/Processos de Negócio/2018

Um desenvolvedor recebeu um conjunto de dados representando o perfil de um grupo de clientes, sem nenhuma informação do tipo de cada cliente, onde cada um era representado por um conjunto fixo de atributos, alguns contínuos, outros discretos. Exemplos desses atributos são: idade, salário e estado civil. Foi pedido a esse desenvolvedor que, segundo a similaridade entre os clientes, dividisse os clientes em grupos, sendo que clientes parecidos deviam ficar no mesmo grupo. Não havia nenhuma informação que pudesse ajudar a verificar se esses grupos estariam corretos ou não nos dados disponíveis para o desenvolvedor.

Esse é um problema de data mining conhecido, cuja solução mais adequada é um algoritmo

a) de regressão

b) não supervisionado

c) por reforço

d) semissupervisionado

e) supervisionado

Comentário: Veja que a questão deu algumas dicas sobre a resposta correta. A principal delas é o fato de o desenvolvedor não ter nenhuma informação do tipo (ou classe) dos clientes. Ele, então, vai fazer uma análise dos dados para construir agrupamentos por similaridades entre os membros de um determinado grupo. Desta forma, podemos perceber que a nossa resposta se encontra na alternativa B.

Gabarito: B

5. ESAF - 2008 - Cargo: ANALISTA DE FINANÇAS E CONTROLE - Secretaria do Tesouro Nacional - STN - TECNOLOGIA DA INFORMAÇÃO/ INFRA-ESTRUTURA DE TI

13- Com respeito à mineração de dados, assinale a opção correta, após avaliar as seguintes afirmações:

I. A mineração de dados pode ser usada em conjunto com um datawarehouse, para auxiliar tomada de decisão.

II. A mineração de dados permite a descoberta de regras de associação entre hierarquias.

III. A mineração de dados compreende todo o processo de descoberta de conhecimento em bancos de dados.

a) Apenas as afirmações I e II são corretas.

b) Apenas as afirmações I e III são corretas.

c) Apenas as afirmações II e III são corretas.

d) As afirmações I, II e III são corretas.

e) As afirmações I, II e III são incorretas.

Comentários: Vamos comentar cada uma das alternativas acima:

I. Exatamente, ambas fazem parte do processo de descoberta de conhecimento. Em um primeiro momento é formado o Data Warehouse com a base de dados que agrega informações de diferentes fontes. Após passar por uma limpeza (para retirar dados inconsistentes, ex.: uma data sem o ano 12/02) esses dados são integrados de forma a compor um DW com informações de todos os setores da organização. Num segundo momento, um algoritmo de mineração de dados toma uma parte dos dados do DW e procura encontrar regras ou padrões úteis.

II. Verdadeiro. Uma das tarefas de Data Mining é a descoberta de regras de associação entre dados. Outras tarefas que podemos listar são:

– Classificação

– Clustering

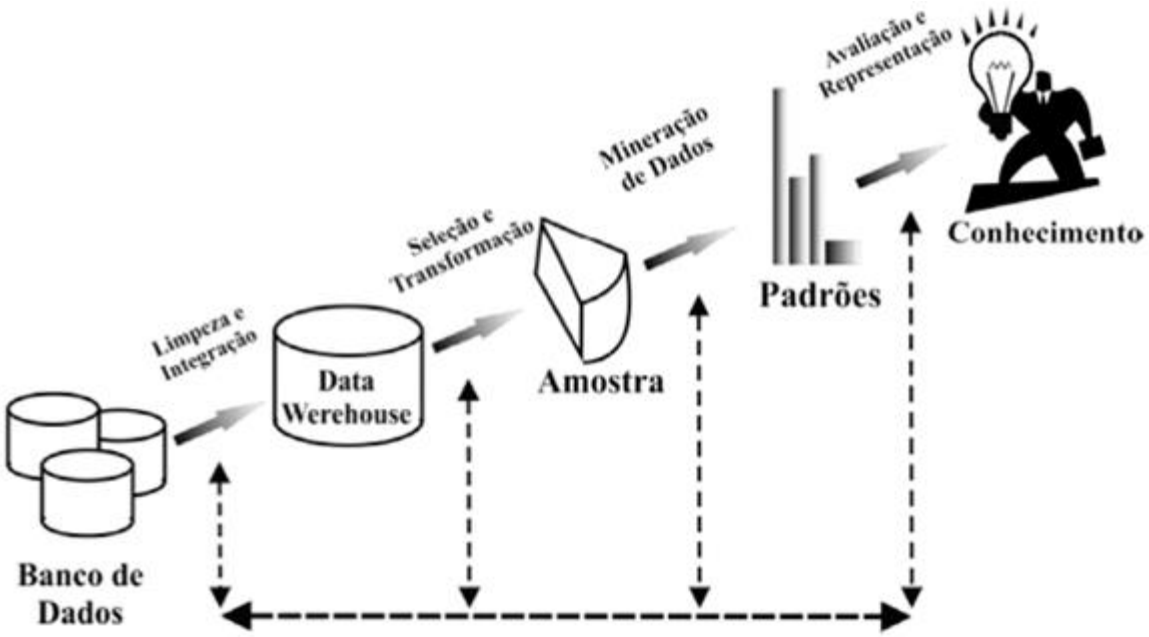
– Estimativa

– Previsão

– Agrupamento por afinidade

III. Na realidade, é o oposto do que está dito na alternativa. É a mineração de dados que faz parte do processo de descoberta do conhecimento.

Para entender melhor, veja a figura abaixo:



Gabarito: A

6. ESAF 2013 – Secretária do Tesouro Nacional – Analista de sistemas

8 - A Mineração de Dados requer uma adequação prévia dos dados através de técnicas de pré-processamento. Entre elas estão as seguintes técnicas:

a) Agrupamento. Amostragem. Redução de dimensionalidade. Seleção de subconjuntos de recursos. Recursos pontuais. Polarização. Redução de variáveis.

b) Agregação. Classificação. Redução de faixas de valores. Seleção de subconjuntos de recursos. Redução de recursos. Terceirização e discretização. Transformação de variáveis.

c) Agrupamento. Classificação. Redução de dimensionalidade. Seleção de subconjuntos de usuários. Criação de recursos. Binarização e discretização. Transformação de conjuntos.

d) Agregação. Amostragem. Redução de dimensionalidade. Seleção de subconjuntos de usuários. Criação de recursos. Polarização. Transformação de conjuntos.

e) Agregação. Amostragem. Redução de dimensionalidade. Seleção de subconjuntos de recursos. Criação de recursos. Binarização e discretização. Transformação de variáveis.

Comentários: Essa questão não mede o grau de conhecimento do candidato e sim a sua capacidade de decorar uma lista, que cada dia cresce mais, de técnicas de data mining, mais precisamente das técnicas de pré-processamento. A lista foi retirada do livro do TAN. Veja abaixo a lista em inglês, uma tradução ao pé da letra encontra-se na alternativa E, na mesma ordem apresentada.

Aggregation

Sampling

Dimensionality Reduction

Feature subset selection

Feature creation

Discretization and Binarization

Attribute Transformation

Gabarito: E

7. ESAF 2013 – Secretária do Tesouro Nacional – Analista de sistemas

Uma técnica de classificação em Mineração de Dados é uma abordagem sistemática para

a) construção de controles de ordenação a partir de um conjunto de acessos.

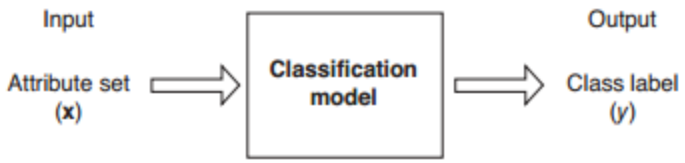
b) construção de modelos de classificação a partir de um conjunto de dados de entrada.

c) construção de modelos de dados a partir de um conjunto de algoritmos.

d) construção de controles de ordenação independentes dos dados de entrada.

e) construção de modelos de sistemas de acesso a partir de um conjunto de algoritmos.

Comentários: Essa questão trata de classificação, uma tarefa de data mining. A classificação pode ser entendida como o processo de encontrar um conjunto de modelos (funções) que descrevem e distinguem classes ou conceitos, com o propósito de utilizar o modelo para predizer a classe de objetos que ainda não foram classificados. Segundo o TAN, a definição também pode ser a tarefa de aprender uma função alvo f que mapeie cada conjunto de atributos x para um dos rótulos de classes y pré-determinados. Com a ajuda das duas definições acima podemos chegar à resposta na letra B. Abaixo, temos uma figura do livro do TAN que apresenta graficamente a definição de classificação.



Gabarito: B

8. ESAF - 2012 - CGU - Analista de Finanças e Controle - prova 3 - Auditoria e Fiscalização - Geral.

São características gerais de conjuntos de dados:

(a) disposição, dispersão e renderização.

(b) dimensão, posicionamento e homogeneidade.

(c) compatibilidade, dispersão e interação.

(d) dimensão, dispersão e resolução.

(e) portabilidade, concentração e resolução.

Comentários: Questão retirada de TAN et. al (2009, p. 35), que destaca três características aplicadas a muitos conjuntos de dados e que possuem um impacto significativo sobre as técnicas de mineração de dados: dimensão, dispersão e resolução.

A **dimensão** refere-se à quantidade de atributos de um conjunto de dados;

A **resolução** está relacionada à granularidade dos dados.

Um conjunto de dados é muito disperso quando para um atributo relevante, a maioria dos valores é NULL ou um valor padrão, e esse conceito está relacionado à **dispersão**.

Gabarito: D

9. ESAF - 2012 - Receita Federal - Analista Tributário da RF - Prova 2 - Área Informática

Um data mining inteligente descobre informações em data warehouses onde consultas e relatórios não conseguem revela-las. Ferramentas dedata mining encontram padrões em dados e podem até deduzir regras a partir deles. Os métodos usados para identificar padrões em dados são:

(a) modelos simples, modelos intermediários e modelos complexos.

(b) modelos simples, modelos físicos e modelos integrados.

(c) modelos híbridos, modelos top-down e modelos bottom-up.

(d) modelos lógicos, modelos físicos e modelos interativos.

(e) modelos básicos, modelos genéricos e modelos complementares.

Comentários: De acordo com TURBAN, são usados três métodos para identificar padrões em dados:

Modelos simples (consultas baseadas em SQL, OLAP, raciocínio humano)

Modelos intermediários (regressão, árvores de decisão, agrupamento)

Modelos complexos (redes neurais, outra indução de regras)

Gabarito: A

10. ESAF - 2012 - CGU - Analista de Finanças e Controle - prova 3 - Auditoria e Fiscalização - Geral

São aspectos motivadores da Mineração de Dados:

(a) Escalabilidade. Dimensionalidade moderada. Dados homogêneos. Propriedade e centralização dos dados.

(b) Extensibilidade. Alta paridade. Dados complexos e heterogêneos. Concorrência e distribuição dos dados.

(c) Escalabilidade. Alta dimensionalidade. Dados complexos e heterogêneos. Propriedade e distribuição de dados.

(d) Escalabilidade. Dimensionalidade variável. Dados compatíveis e acoplados. Adequação da distribuição de dados.

(e) Especialidade. Alta dimensionalidade de verificação. Dados complexos e complementares. Propriedade e consistência de dados.

Comentários: A banca retirou os aspectos de TAN e todos os que foram listados na assertiva C estão corretos. Um aspecto abordado no livro e não mencionado na questão é o de “Análises não tradicionais”.

Gabarito: C

11. ESAF - 2012 - CGU - Analista de Finanças e Controle - prova 3 - Auditoria e Fiscalização - Geral

Classificação é

(a) a tarefa de atualizar uma função focal f que permeia cada conjunto de variáveis x para um dos blocos de classes y discretos.

(b) o mapeamento de uma função objetivo f à qual são atribuídos valores x fixados por categorias de rótulos de classes z pré-determinados.

(c) a função alvo f que mapeie cada classificação de atributos x para um dos eixos de classes y pré-determinados.

(d) a tarefa de aprender uma função alvo f que mapeie cada conjunto de atributos x para um dos rótulos de classes y pré-determinados.

(e) a tarefa de ordenar funções de mapeamento para cada categoria de atributos x para um dos rótulos de variáveis y controladas.

Comentários: Como vimos ao longo da nossa aula, a classificação é a tarefa de aprendizado de uma função alvo f que mapeia cada atributo de um conjunto x para um rótulo de classe predefinido y. Essa definição foi retirada do livro do TAN que inclusive possui o arquivo em pdf do capítulo de classificação disponível na web.

Gabarito: D

12. ESAF - 2012 - CGU - Analista de Finanças e Controle - prova 3 - Auditoria e Fiscalização - Geral

A Mineração de Dados é

(a) o processo de desenvolvimento de soluções automáticas de acesso a informações úteis em depósitos de dados.

(b) a transformação automática de dados existentes em grandes depósitos de dados em informações quantificáveis.

(c) a automação da recuperação de informações caracterizadas por registros com grande quantidade de atributos.

(d) a descoberta de relações significativas entre dados e informações passíveis de atualização automática.

(e) o processo de descoberta automática de informações úteis em grandes depósitos de dados.

Comentários: Essa questão peca por falta de preciosismo do examinador. Sabemos que o processo de mineração de dados requer supervisão. Ele não é totalmente automático, embora várias etapas desse processo sejam automatizáveis. Mas, se analisarmos cada uma das alternativas acima, podemos observar que aquela que mais se aproxima de todas as definições que vimos até aqui é a presente na alternativa E.

Gabarito: E

13. ESAF – CVM 2010 - Sistemas.

53- Mineração de Dados é

a) o processo de atualizar de maneira semiautomática grandes bancos de dados para encontrar versões úteis.

b) o processo de analisar de maneira semiautomática grandes bancos de dados para encontrar padrões úteis.

c) o processo de segmentar de maneira semiautomática bancos de dados qualitativos e corrigir padrões de especificação.

d) o programa que depura de maneira automática bancos de dados corporativos para mostrar padrões de análise.

e) o processo de automatizar a definição de bancos de dados de médio porte de maior utilidade para os usuários externos de rotinas de mineração.

Comentários: Vejam que esta questão, quando contrastada com a anterior, leva-nos a uma visão mais precisa do conceito de mineração de dados. Em ambos os casos, o objetivo final é descobrir informações úteis, mas no caso desta questão o examinador se preocupou em deixar claro que é um processo semiautomático.

Gabarito: B

14. ANO: 2015 BANCA: FCC ÓRGÃO: CNMP PROVA: ANALISTA DO CNMP - DESENVOLVIMENTO DE SISTEMAS

Em relação às ferramentas de Data Discovery e os fundamentos de Data Mining, é correto afirmar:

A Data Mining é o processo de descobrir conhecimento em banco de dados, que envolve várias etapas. O KDD – Knowledge Discovery in Database é uma destas etapas, portanto, a mineração de dados é um conceito que abrange o KDD.

B A etapa de KDD do Data Mining consiste em aplicar técnicas que auxiliem na busca de relações entre os dados. De forma geral, existem três tipos de técnicas: Estatísticas, Exploratórias e Intuitivas. Todas são devidamente experimentadas e validadas para o processo de mineração.

C Os dados podem ser não estruturados (bancos de dados, CRM, ERP), estruturados (texto, documentos, arquivos, mídias sociais, cloud) ou uma mistura de ambos (emails, SOA/web services, RSS). As ferramentas de Data Discovery mais completas possuem conectividade para todas essas origens de dados de forma segura e controlada.

D Estima-se que, atualmente, em média, 80% de todos os dados disponíveis são do tipo estruturado. Existem diversas ferramentas open source e comerciais de Data Discovery. Dentre as open source está a InfoSphere Data Explorer e entre as comerciais está a Vivisimo da IBM.

E As ferramentas de Data Mining permitem ao usuário avaliar tendências e padrões não conhecidos entre os dados. Esses tipos de ferramentas podem utilizar técnicas avançadas de computação como redes neurais, algoritmos genéticos e lógica nebulosa, dentre outras.

Comentário: Vamos fazer alguns comentários interessantes sobre essa questão. Primeiramente, Mineração de Dados é parte de um processo maior de pesquisa denominado Busca de Conhecimento em Banco de Dados (*Knowledge Discovery in*

Database - KDD), o qual possui uma metodologia própria para preparação e exploração dos dados, interpretação de seus resultados e assimilação dos conhecimentos minerados. No entanto, tornou-se mais conhecida do que o próprio processo de KDD em função de ser a etapa onde são aplicadas as técnicas de busca de conhecimentos.

Os métodos de data mining são tecnologias existentes, independente do contexto mineração de dados, uma vez que, aplicados na KDD, produzem bons resultados, transformando dados em conhecimento útil e favorecendo as práticas de estudos baseados em evidências. São vários métodos existentes e utilizados, entre eles, temos: Rede Neurais, Árvore de Decisão, Algoritmos Genéticos (AGs), Lógica Nebulosa (*Fuzzy logic*) e Estatística.

Um método muito utilizado é a Lógica Nebulosa (*Fuzzy logic*), uma teoria matemática que permite uma modelagem do modo aproximado de raciocínio, imitando a habilidade humana de tomar decisões em ambientes de incertezas e imprecisão. Com isso, pode-se construir sistemas inteligentes de controle e suporte à decisão.

Analisando a explicação teórica, podemos perceber que a alternativa E se encontra correta. Como exercício, você pode encontrar os erros das demais alternativas.

Gabarito: E.

15. ANO: 2014 BANCA: FCC ÓRGÃO: TCE-RS PROVA: AUDITOR PÚBLICO EXTERNO - TÉCNICO EM PROCESSAMENTO DE DADOS

A revista da CGU − Controladoria Geral da União, em sua 8a edição, publicou um artigo que relata que foram aplicadas técnicas de exploração de dados, visando a descoberta de conhecimento útil para auditoria, em uma base de licitações extraída do sistema ComprasNet, em que são realizados os pregões eletrônicos do Governo Federal. Dentre as técnicas preditivas e descritivas utilizadas, estão a classificação, clusterização e regras de associação. Como resultado, grupos de empresas foram detectados em que a média de participações juntas e as vitórias em licitações levavam a indícios de conluio.

As técnicas aplicadas referem-se a

A Customer Churn Trend Analysis.

B On-Line Analytical Processing.

C Data Mining.

D Business Process Management.

E Extraction, Transformation and Load.

Comentário: Vejam que todas as técnicas se referem às atividades de data mining. Vamos rapidamente relembrar o conceito de cada uma delas:

É possível usar a **análise preditiva** para resolver seus problemas mais difíceis. Ela ajuda a descobrir padrões no passado que podem sinalizar o que está por vir. Essa análise é capaz de descobrir padrões ocultos nos dados que o especialista humano pode não ver. Ela é, na verdade, o resultado de matemática e estatística aplicada aos dados.

As **técnicas descritivas** ou **exploratórias** são utilizadas para organizar os dados e investigá-los, relatar ou expor suas características e procurar indícios de padrões ou características interessantes que possam indicar possíveis tendências.

A classificação representa a generalidade de problemas de mineração de dados atualmente, por meio da criação de modelos de classes para um conjunto de objetos. Após a definição de um conjunto de classes, novos objetos cadastrados na base de dados podem ser classificados de acordo com as classes previamente definidas.

As **técnicas de clusterização** procuram semelhanças e diferenças num conjunto de dados e agrupam os registos semelhantes em segmentos ou clusters, de uma forma automática, de acordo com algum critério ou métrica.

A **associação** visa solucionar problemas de análise de cesta de produtos, gerando modelos descritivos que permitem descobrir regras ou padrões de consumo de clientes.

Gabarito: C.

16. ANO: 2014 BANCA: FCC ÓRGÃO: TRF 3ª REGIÃO (SP MS) PROVA: ANALISTA JUDICIÁRIO - INFORMÁTICA (BANCO DE DADOS)

Mineração de dados é a investigação de relações e padrões globais que existem em grandes bancos de dados, mas que estão ocultos no grande volume de dados. Com base nas funções que executam, há diferentes técnicas para a mineração de dados, dentre as quais estão:

I. identificar afinidades existentes entre um conjunto de itens em um dado grupo de registros. Por exemplo: 75% dos envolvidos em processos judiciais ligados a ataques maliciosos a servidores de dados também estão envolvidos em processos ligados a roubo de dados sigilosos.

II. identificar sequências que ocorrem em determinados registros. Por exemplo: 32% de pessoas do sexo feminino após ajuizarem uma causa contra o INSS solicitando nova perícia médica ajuízam uma causa contra o INSS solicitando ressarcimento monetário.

III. as categorias são definidas antes da análise dos dados. Pode ser utilizada para identificar os atributos de um determinado grupo que fazem a discriminação entre 3 tipos diferentes, por exemplo, os tipos de processos judiciais podem ser categorizados como infrequentes, ocasionais e frequentes.

Os tipos de técnicas referenciados em I, II e III, respectivamente, são:

A Padrões sequenciais - Redes Neurais - Árvore de decisão

B Redes Neurais - Árvore de decisão - Padrões sequenciais

C Associação - Padrões sequenciais - Classificação

D Classificação - Associação - Previsão

E Árvore de decisão - Classificação – Associação

Comentário: Já descrevemos as técnicas utilizadas em mineração de dados que podem ser utilizadas em diferentes contextos. Agora, podemos associar o item I às regras de associação, o item 2 a padrões sequenciais e o item III à classificação. Assim, temos o gabarito na alternativa C.

Gabarito: C.

17. ANO: 2013 BANCA: FCC ÓRGÃO: MPE-MA PROVA: ANALISTA JUDICIÁRIO - BANCO DE DADOS

Uma das funções desempenhadas pelas técnicas de mineração de dados consiste em determinar que itens de um conjunto de dados ocorrem de forma simultânea. Essa função recebe a denominação de

A análise de afinidade.

B estimativa.

C previsão.

D seleção adaptativa.

E análise de variância.

Comentário: Vamos começar definindo cada alguns dos termos acima que considero relevantes:

Análise de Afinidade – Como o nome já diz, essa técnica determina que alguns fatos ocorrem simultaneamente com probabilidade razoável, ou então que itens de dados estão presentes conjuntamente com uma chance razoável. Um exemplo disso é o de um carrinho de supermercado, por meio dele pode-se extrair informações para que a organização dos produtos no supermercado agrade aos consumidores, colocando próximos uns aos outros produtos comprados em conjunto.

Estimativa – Esta técnica é utilizada para determinar um valor aproximado de uma variável por meio de dados que foram passados ou de dados adquiridos de outras variáveis semelhantes, sobre os quais se tem conhecimento.
Previsão – Esta técnica tem por objetivo a avaliação de um valor de uma variável ainda não identificada, baseando-se em dados adquiridos por meio do comportamento desta variável ao longo do tempo.

Vejam que, analisando as descrições dadas para cada um dos termos presentes nas alternativas, podemos concluir que a resposta se encontra na alternativa A.

Gabarito: A.

18. ANO: 2010 BANCA: FCC ÓRGÃO: TCE-SP PROVA: AGENTE DA FISCALIZAÇÃO FINANCEIRA - PRODUÇÃO E BANCO DE DADOS

NÃO é um objetivo da mineração de dados (mining), na visão dos diversos autores,

A garantir a não redundância nos bancos transacionais.

B conhecer o comportamento de certos atributos no futuro.

C possibilitar a análise de determinados padrões de eventos.

D categorizar perfis individuais ou coletivos de interesse comercial.

E apoiar a otimização do uso de recursos limitados e/ou maximizar variáveis de resultado para a empresa.

Comentário: Essa questão serve para aprendermos um pouco sobre os **objetivos de mineração de dados**. Vejam que a resposta está na alternativa A, sabemos que a garantia de não redundância dos bancos de dados transacionais **não** tem nenhuma relação com um modelo multidimensional usado para análise de dados.

Por outro lado, observem as demais alternativas. Cada uma delas apresente um dos possíveis objetivos das tarefas de data mining.

Gabarito: A

19. ANO: 2010 BANCA: FCC ÓRGÃO: TCE-SP PROVA: AGENTE DA FISCALIZAÇÃO FINANCEIRA - PRODUÇÃO E BANCO DE DADOS

A data mining apoia o conhecimento indutivo que pode ser representado por

I. Lógica proposicional.

II. Árvores de decisão.

III. Redes neurais.

IV. Redes semânticas.

Está correto o que consta em

A I e III, apenas.

B II e III, apenas.

C II, III e IV, apenas.

D I, II e IV, apenas.

E I, II, III e IV.

Comentário: Segundo o Navathe, Data Mining apoia o conhecimento indutivo, que descobre novas regras e padrões nos dados fornecidos. O conhecimento pode ser representado de muitas formas. Em um senso não estruturado, pode ser representado por **regras ou por lógica proposicional**. Em uma forma estruturada, pode ser representado por **árvores de decisão, redes semânticas, redes neurais** ou hierarquias de classes ou frames. Desta forma, podemos concluir que todas as alternativas estão corretas e nossa resposta encontra-se na alternativa E.

Gabarito: E

20. ANO: 2010 BANCA: FCC ÓRGÃO: TCE-SP PROVA: AGENTE DA FISCALIZAÇÃO FINANCEIRA - PRODUÇÃO E BANCO DE DADOS

No âmbito dos algoritmos associados ao mining, se houver um banco de dados com um número potencial pequeno de conjuntos de itens grandes, isto é, uns poucos milhares, então o suporte para todos eles pode ser testado em uma passagem usando a técnica específica de

A hierarquização.

B partição.

C amostragem.

D árvore de padrão frequente.

E séries temporais.

Comentário: Essa foi mais uma questão que foi retirada do livro do Navathe. Vamos analisar cada uma das alternativas para entender cada um dos conceitos:

Hierarquização: Existem alguns tipos de associações que são particularmente interessantes por alguma razão especial. Essas associações ocorrem entre hierarquias de itens. Tipicamente, é possível dividir itens entre hierarquias separadas baseadas na natureza do domínio. Após dividir os itens em hierarquias distintas, nosso interesse se concentra nas descobertas de regras de associações entre as diferentes hierarquias.

Partição: Se tivermos um banco de dados com um número potencial pequeno de conjuntos de itens grandes, digamos, uns poucos milhares, então o suporte para todos eles podem ser testados em uma passagem usando a **técnica de partição**.

Amostragem: A ideia principal do **algoritmo de amostragem** é selecionar uma pequena amostra, que caiba na memória principal do banco de dados de transações, e determinar os conjuntos de itens frequentes daquela amostra.

Árvore de padrão frequente: O **algoritmo de árvore padrão frequente** é motivado pelo fato de que os algoritmos baseados no algoritmo Apriori podem gerar e testar um número muito grande de conjunto de itens candidatos.

Séries temporais: são sequências de eventos, cada evento pode ser um tipo fixo dado uma transação. Por exemplo, o preço de fechamento de uma ação na bolsa de valores.

Desta forma, podemos encontrar nossa resposta correta na alternativa B.

Gabarito: B

21. ANO: 2010 BANCA: FCC ÓRGÃO: TCE-SP PROVA: AGENTE DA FISCALIZAÇÃO FINANCEIRA - PRODUÇÃO E BANCO DE DADOS

Uma das abordagens de mining define que, se uma regra de classificação é considerada uma função sobre variáveis que as mapeia em uma classe destino, a regra é chamada

A categorização.

B Apriori.

C algoritmo genético.

D regressão.

E minimização.

Comentário: Mais uma questão da FCC cuja referência é o livro do Navathe, vejamos o que ele tem a dizer sobre regressão:

Regressão é uma aplicação especial da regra de classificação. Se uma regra de classificação é considerada uma função sobre variáveis que as mapeia em uma classe destino, a regra é chamada regressão. Uma aplicação de regressão ocorre quando, em vez de mapear uma tupla de dados de uma relação para uma classe específica, o valor da variável é previsto baseado naquela tupla.

Outro termo que é comentado na questão e que ainda não falamos são os algoritmos genéticos. **Algoritmos Genéricos** ou AGs são algoritmos de otimização e busca baseados nos mecanismos de seleção natural e genética. Enquanto os métodos de otimização e busca convencionais trabalham geralmente de forma sequencial, avaliando a cada instante uma possível solução, os AGs trabalham com um conjunto de possíveis soluções simultaneamente. Algoritmos Genéticos (AGs) são uma classe de procedimentos de pesquisa aleatórios capazes de realizar pesquisas adaptativas e robustas sobre uma ampla gama de topologias de espaço de pesquisa. Modelados após o surgimento adaptativo de espécies biológicas a partir de mecanismos evolutivos e introduzidos por Holland, AGs vêm sendo aplicados com sucesso em campos diversificados como análise de imagens, escalonamentos e projetos de engenharia.

Vejam que, pelo exposto acima, podemos marcar nossa resposta na **alternativa D**.

Gabarito: D

22. ANO: 2010 BANCA: FCC ÓRGÃO: TCE-SP PROVA: AGENTE DA FISCALIZAÇÃO FINANCEIRA - PRODUÇÃO E BANCO DE DADOS

Considere uma dada população de eventos ou novos itens que podem ser particionados (segmentados) em conjuntos de elementos similares, tal como, por exemplo, uma população de dados sobre uma doença que pode ser dividida em grupos baseados na similaridade dos efeitos colaterias produzidos. Como um dos modos de descrever o conhecimento descoberto durante a data mining este é chamado de

A associação.

B otimização.

C classificação.

D clustering.

E temporização.

Comentário. Vamos voltar para as definições vistas ao longo do texto teórico da aula. Essa questão vai nos ajudar a fixar conceitos fazendo uma revisão sobre o assunto. Faremos uma definição de cada um dos termos listados nas alternativas, exceto de temporização que não faz parte do escopo de mineração.

Associação: Devido a sua grande aplicabilidade, as **regras de associação** encontram-se entre um dos mais importantes tipos de conhecimento que podem ser minerados em bases de dados. Estas regras representam padrões de relacionamento entre itens de uma base de dados. Uma de suas típicas aplicações é a **análise de transações de compras** (*market basket analysis*), um processo que examina padrões de compras de consumidores para determinar produtos que costumam ser adquiridos em conjunto. Um exemplo de regra de associação, obtida a partir da análise de uma base de dados real, que registra os produtos adquiridos por famílias cariocas em suas compras mensais, é dado por: {*mini-pizza semi-pronta*} Þ {*suco de fruta em pó*}. Esta regra de associação indica que as famílias que compram o produto {*mini-pizza semi-pronta*} tem **maior chance** de também adquirir o produto {*suco de fruta em pó*}.

Otimização: Esta funcionalidade visa otimizar recursos limitados como tempo, espaço, dinheiro, matéria-prima etc., buscando maximizar variáveis de resultado como vendas, lucros, distribuição, economia de espaço etc. Esta funcionalidade se aproxima dos estudos da área de pesquisa operacional, a qual trata de problemas de otimização, sempre sujeito a um conjunto de restrições. Como exemplo, podemos estudar as vendas de um supermercado, no sentido de otimizar a distribuição de seus produtos em suas gôndolas, visando otimizar a exposição de um número cada vez maior de produtos.

Classificação: A classificação consiste em examinar uma certa característica nos dados e atribuir uma classe previamente definida. Dados podem ser associados a classes ou a conceitos por meio de um processo de discriminação ou de caracterização. Discriminação se caracteriza por ter seu resultado obtido por meio da atribuição de um valor a um atributo no registro, em função de um ou mais de seus atributos. Por exemplo, em um supermercado podemos classificar os produtos por tipo como alimentício, vestuário, higiene e limpeza etc. Já caracterização é a sumarização de um atributo de estudo por uma característica de um ou mais atributos. Por exemplo, podemos caracterizar um empregado pelo seu salário anual, identificando faixas da agregação mensal de seus salários em baixa, média e alta.

Clustering: Esta funcionalidade visa segmentar um conjunto de dados num número de **subgrupos homogêneos** ou ***clustering***. Seu objetivo é formar grupos baseados no princípio de que esses grupos devem ser o mais homogêneo em si e mais heterogêneo entre si. A diferença fundamental entre a formação de agrupamento e a classificação é que no agrupamento não existem classes predefinidas para classificar os registros em estudo. Os registros são agrupados em função de suas similaridades básicas, ou seja, quando se deseja formar agrupamentos, seleciona-se um conjunto de atributos (variáveis) e em função da similaridade desses atributos são formados os grupos.

Analisando cada um dos termos acima, podemos concluir que nossa resposta está na **alternativa D**.

Gabarito: D

23. ANO: 2015 BANCA: FGV ÓRGÃO: TJ-SC PROVA: ANALISTA JUDICIÁRIO - ANALISTA DE SISTEMAS

João trabalha no setor de BI da empresa e recebeu a tarefa de identificar agrupamentos de alunos de uma escola segundo seu desempenho acadêmico. A partir das notas obtidas, João deve formar grupos tal que integrantes de um grupo tenham desempenho similar, e que integrantes de grupos distintos sejam dissimilares. O algoritmo mais apropriado para essa tarefa é:

A Apriori;

B decision tree;

C PageRank;

D CART;

E k-means.

Comentário. Essa questão apresenta algoritmos específicos de mineração. Eles são usados na implementação prática das tarefas de mineração. Vamos entender o que cada um deles faz.

Algoritmo Apriori: Existem alguns algoritmos para a geração de **regras de associação**. Talvez o mais popular seja o algoritmo **apriori**, implementado em diversas ferramentas de Data Mining (mineração de dados), como o Weka. Este algoritmo recebe como parâmetro um conjunto de transações T, o valor percentual S como o **suporte** e um valor percentual C para a **confiança**. O algoritmo gera um conjunto de regras no formato A => B [suporte, confiança], em que o conjunto A é chamado de **antecedente** da regra e o conjunto B é chamado de **consequente**. Cada regra gerada deve ter seu suporte e sua confiança maior ou igual ao suporte e confiança mínimo passado para o algoritmo, respectivamente.

Árvore de decisão: As árvores de decisão constituem uma técnica muito poderosa e amplamente utilizada em **problemas de classificação**. Uma das razões para que esta técnica seja bastante utilizada é o fato do conhecimento adquirido ser representado por meio de regras. Essas regras podem ser expressas em linguagem natural, facilitando o entendimento por parte das pessoas.

PageRank: PageRank™ é um algoritmo utilizado pela ferramenta de busca Google para posicionar websites entre os resultados de suas buscas. O PageRank mede a importância de uma página contabilizando a quantidade e qualidade de links apontando para ela. Não é o único algoritmo utilizado pelo Google para classificar páginas da internet, mas é o primeiro utilizado pela companhia e o mais conhecido.

CART: CART significa *Classification and Regression Trees*. Uma árvore de classificação utilizando a metodologia CART traduz o resultado de uma partição binária recursiva dos dados base da modelagem. Este algoritmo é um modelo de regressão não paramétrico que estabelece uma relação entre as variáveis independentes (x), com uma única variável dependente, ou resposta, (*target*) ou alvo. O modelo é ajustado mediante sucessivas divisões binárias no conjunto de dados, para tornar os subconjuntos de dados da variável resposta cada vez mais homogêneos.

k-means: Proposto por J. MacQueen em 1967, o algoritmo de **Análise de Agrupamento k-means** (ou *algoritmo das k-médias*) é um dos mais conhecidos e utilizados, além de ser o que possui o maior número de variações. O algoritmo inicia com a escolha dos k elementos que formaram as sementes iniciais. Esta escolha pode ser feita de muitas formas, entre elas:

Selecionando as k primeiras observações;

Selecionando k observações aleatoriamente; e

Escolhendo k observações de modo que seus valores sejam bastante diferentes. Por exemplo, ao se agrupar uma população em três grupos de acordo com a altura dos indivíduos, poderia se escolher um indivíduo de baixa estatura, um de estatura mediana e um alto.

Escolhidas as sementes iniciais, é calculada a distância de cada elemento em relação às sementes, agrupando o elemento ao grupo que possuir a menor distância (mais similar) e recalculando o seu centroide. O processo é repetido até que todos os elementos façam parte de um dos clusters. Vejam que esse é exatamente o objetivo do João! Desta forma, nossa resposta encontra-se na **alternativa E**.

Gabarito: E

CONSIDERAÇÕES FINAIS

Chegamos ao final da nossa aula que abordou os assuntos relacionados à Data Mining e Aprendizado de Máquina.

Esperamos que você tenha gostado e aprendido bastante sobre o assunto.

Até a próxima!

Thiago Cavalcanti

ESSA LEI TODO MUNDO CONHECE:
PIRATARIA É CRIME.

Mas é sempre bom revisar o porquê e como você pode ser prejudicado com essa prática.

1. 

Professor investe seu tempo para elaborar os cursos e o site os coloca à venda.

2. 

Pirata divulga ilicitamente (grupos de rateio), utilizando-se do anonimato, nomes falsos ou laranjas (geralmente o pirata se anuncia como formador de "grupos solidários" de rateio que não visam lucro).

3. 

Pirata cria alunos fake praticando falsidade ideológica, comprando cursos do site em nome de pessoas aleatórias (usando nome, CPF, endereço e telefone de terceiros sem autorização).

4. 

Pirata compra, muitas vezes, clonando cartões de crédito (por vezes o sistema anti-fraude não consegue identificar o golpe a tempo).

5. 

Pirata fere os Termos de Uso, adultera as aulas e retira a identificação dos arquivos PDF (justamente porque a atividade é ilegal e ele não quer que seus fakes sejam identificados).

6. 

Pirata revende as aulas protegidas por direitos autorais, praticando concorrência desleal e em flagrante desrespeito à Lei de Direitos Autorais (Lei 9.610/98).

7. 

Concurseiro(a) desinformado participa de rateio, achando que nada disso está acontecendo e esperando se tornar servidor público para exigir o cumprimento das leis.

8. 

O professor que elaborou o curso não ganha nada, o site não recebe nada, e a pessoa que praticou todos os ilícitos anteriores (pirata) fica com o lucro.

Deixando de lado esse mar de sujeira, aproveitamos para agradecer a todos que adquirem os cursos honestamente e permitem que o site continue existindo.